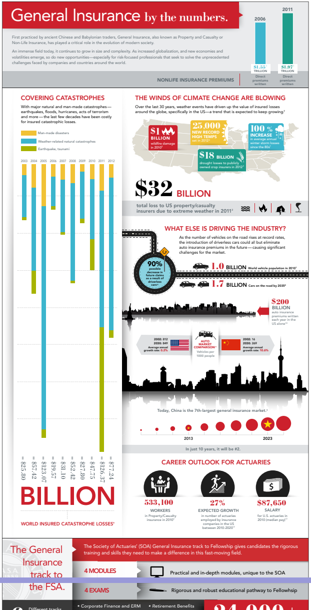
Today

- data presentation Michael
- In the news
- Semi-parametric regression

- HW 3: due March 21

- Final Exam: April 11 2:00 – 5:00 pm SS 1085
  - 4 questions
  - one theory question
  - one applied question
  - one question from HW
  - one question about a study
  - one question with computer output
  - detailed list of SM sections coming
  - See Final Exam from 2012 on course web page

link


General Insurance by the numbers.
2006
2011
$1.55 TRILLION Direct premiums written
$1.97 TRILLION Direct premiums written
First practiced by ancient Chinese and Babylonian traders, General Insurance, also known as Property and Casualty or Non-Life Insurance, has played a critical role in the evolution of modern society.
An immense field today, it continues to grow in size and complexity. As increased globalization, and new economies and volatilities emerge, so do new opportunities—especially for risk-focused professionals that seek to solve the unprecedented challenges faced by companies and countries around the world.
NONLIFE INSURANCE PREMIUMS
COVERING CATASTROPHES
With major natural and man-made catastrophes—earthquakes, floods, hurricanes, acts of terrorism and more—the last few decades have been costly for insured catastrophic losses.
Man-made disasters
Weather-related natural catastrophes
Earthquake, tsunami
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
=$25.80 =$57.42 =$123.07 =$19.57 =$31.10 =$52.42 =$27.80 =$47.75 =$126.37 =$77.24
BILLION
WORLD INSURED CATASTROPHE LOSSES
THE WINDS OF CLIMATE CHANGE ARE BLOWING
Over the last 30 years, weather events have driven up the value of insured losses around the globe, specifically in the US—a trend that is expected to keep growing.
$1 BILLION wildfire damage in 2010
25,000 NEW RECORD HIGH TEMPS set in 2012
100% INCREASE in average annual winter storm losses since the 80s
$18 BILLION drought losses to publicly owned crop insurers in 2012
$32 BILLION
total loss to US property/casualty insurers due to extreme weather in 2011
WHAT ELSE IS DRIVING THE INDUSTRY?
As the number of vehicles on the road rises at record rates, the introduction of driverless cars could all but eliminate auto insurance premiums in the future—causing significant challenges for the market.
90% possible decrease in future claims as a result of driverless cars
1.0 BILLION World vehicle population in 2010
1.7 BILLION Cars on the road by 2035
$200 BILLION auto insurance premiums written each year in the US alone
AUTO MARKET COMPARISON
Vehicles per 1000 people
Today, China is the 7th-largest general insurance market.
2013
2023
In just 10 years, it will be #2.
CAREER OUTLOOK FOR ACTUARIES
533,100 WORKERS in Property/Casualty insurance in 2010
27% EXPECTED GROWTH in number of actuaries employed by insurance companies in the US between 2010-2020
$87,650 SALARY for U.S. actuaries in 2010 (median pay)
The General Insurance track to the FSA.
The Society of Actuaries' (SOA) General Insurance track to Fellowship gives candidates the rigorous training and skills they need to make a difference in this fast-moving field.
4 MODULES
Practical and in-depth modules, unique to the SOA
4 EXAMS
Rigorous and robust educational pathway to Fellowship
• Corporate Finance and ERM
• Retirement Benefits

CBC News

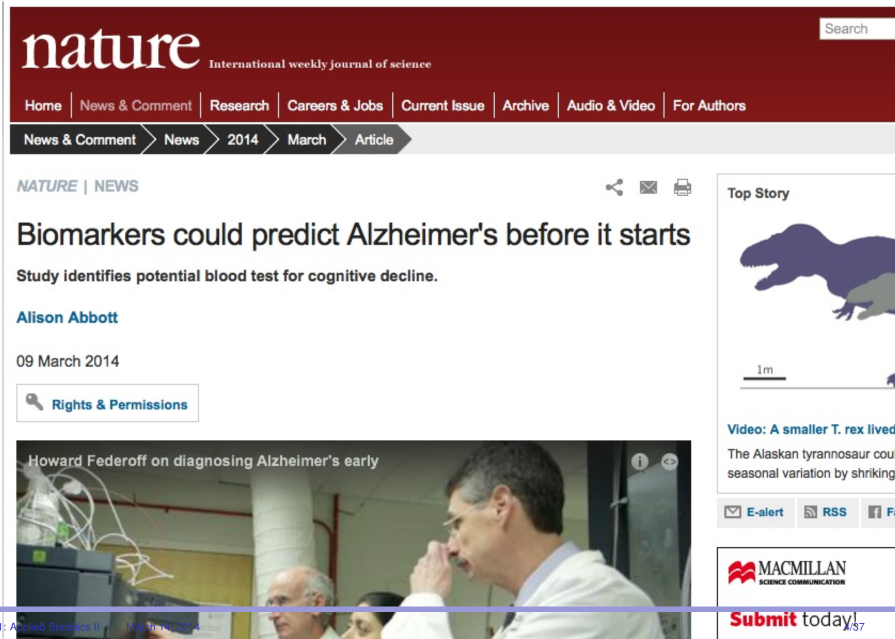
Can a new test to identify the likelihood of Alzheimer's lead to better treatment?

Wednesday, March 12, 2014 | Categories: **Episodes** 0

Like 244 Tweet 42 +1 7 Share 286

Researchers at Georgetown University announced the discovery of a blood test that can predict whether someone will develop Alzheimer's disease. REUTERS / Luis Galdamez

Nature Medicine News Release



nature International weekly journal of science

Search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2014 > March > Article

NATURE | NEWS

Biomarkers could predict Alzheimer's before it starts

Study identifies potential blood test for cognitive decline.

Alison Abbott

09 March 2014

Rights & Permissions

Howard Federoff on diagnosing Alzheimer's early

Top Story

1m

Video: A smaller T. rex lived

The Alaskan tyrannosaur cou
seasonal variation by shriking

E-alert RSS F

MACMILLAN SCIENCE COMMUNICATION

Submit today!

Nature Medicine Advance Publication

nature medicine

Access provided to University of Toro

Search

nature.com ▸ journal home ▸ advance online publication ▸ letter ▸ full text

NATURE MEDICINE | LETTER

Plasma phospholipids identify antecedent memory impairment in older adults

Mark Mapstone, Amrita K Cheema, Massimo S Fiandaca, Xiaogang Zhong, Timothy R Mhyre, Linda H MacArthur, William J Hall, Susan G Fisher, Derick R Peterson, James M Haley, Michael D Nazar, Steven A Rich, Dan J Berlau, Carrie B Peltz, Ming T Tan, Claudia H Kawas & Howard J Federoff

Affiliations | Contributions | Corresponding author

Nature Medicine (2014) | doi:10.1038/nm.3466

Received 27 August 2013 | Accepted 09 January 2014 | Published online 09 March 2014

PDF | Citation | Reprints | Rights & permissions | Article metrics

Alzheimer's disease causes a progressive dementia that currently affects over 35 million

Nature Medicine

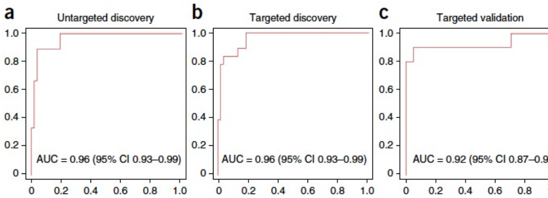
- 525 patients, followed for five years
- 46 patients had AD or pre-AD("aMCI") at entry; 28 converters
- in year 3 53 patients with aMCI/AD selected for plasma testing
- matched with 53 normal controls

- looked for biomarkers of disease, using logistic regression and lasso
- used the most promising to test on a further set of 21 patients with 20 matched controls

- ROC curve: plot of sensitivity (True Positives) against 1-specificity (False Positives) as cut-off varies

... nature medicine

Figure 2 ROC results for the lipidomics analyses. (**a**–**c**) Plots of ROC results from the models derived from the three phases of the lipidomics analysis. Simple logistic models using only the metabolites identified in each phase of the lipidomics analysis were developed and applied to determine the success of the models for classifying the C_{pre} and NC groups. The red line in each plot represents the AUC obtained from the discovery-phase LASSO analysis (**a**), the targeted analysis of the ten metabolites in the discovery phase (**b**) and the application of the ten-metabolite panel developed from the targeted discovery phase in the independent validation phase (**c**). The ROC plots represent sensitivity (i.e., true positive rate) versus 1 – specificity (i.e., false positive rate).



Lasso for choosing explanatory variables

- penalized least squares

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

- equivalent to

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

- quadratic programming problem
- $\hat{\beta}^{lasso}$ is nonlinear function of y

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

- `http://www-stat.stanford.edu/~tibs/lasso.html`
- extends to generalized linear models, implemented in `glmnet`

Extensions of semi-parametric regression

- original model $y_j = g(x_j) + \epsilon_j$

- fit by local polynomial regression:
  $g(x_0) \doteq \beta_0 + \beta_1(x_j - x_0) + \cdots + \beta_k(x_j - x_0)^k, \quad \hat{g}(x_0) = \hat{\beta}_0$
- $\hat{\beta}$ maximizes

$$\ell(\beta, \sigma; x_0, h) = \sum \frac{1}{h} w\left(\frac{x_j - x_0}{h}\right) \ell_j(\beta, \sigma; x_0)$$

- $\ell_j(\beta, \sigma; x_0) = -\frac{1}{2\sigma^2}\{y_j - \beta_0 - \beta_1(x_j - x_0) - \cdots - \beta_k(x_j - x_0)^k\}^2$
  $-\frac{1}{2}\log \sigma^2$
- local log-likelihood fitting

- extend to more general models by replacing ℓ_j by the appropriate log-likelihood contribution

Example 10.32

- toxoplasmosis data; response – incidence;
  x – yearly rainfall SM Figure 10.12
- $y_j = r_j/m_j, \quad r_j \sim \text{Binom}\{m,\pi(x_j)\}$

- $\pi(x) = \exp[\theta(x)/\{1+\exp\{\theta(x)\}]$
- $\theta(x) \doteq \beta_0 + \beta_1(x-x_0) + \cdots + \beta_k(x-x_0)^k/k!, \quad \hat{\theta}(x_0) = \hat{\beta}_0$
- local log-likelihood

$$\ell(\beta; x_0, h) = \sum \frac{1}{h} w\left(\frac{x_j - x_0}{h}\right) m_j\{y_j x_j^{\mathrm{T}}\beta - \log(1+e^{x_j^{\mathrm{T}}\beta})\}$$

- or possibly allow for over-dispersion

$$\ell(\beta, \phi; x_0, h) = \sum \frac{1}{h} w\left(\frac{x_j - x_0}{h}\right) \frac{m_j}{\phi}\{y_j x_j^{\mathrm{T}}\beta - \log(1+e^{x_j^{\mathrm{T}}\beta})\}$$

Example 10.32

528 10 · *Nonlinear Regression Models*

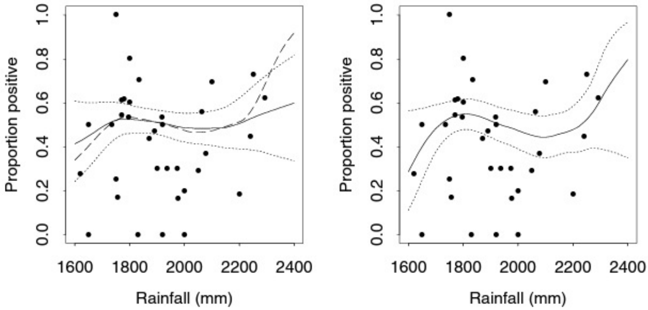
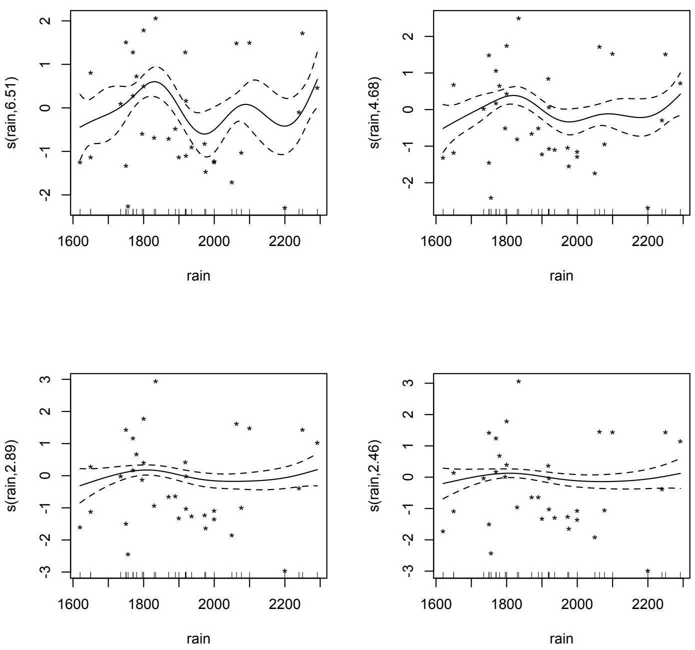


Figure 10.17 Local fits to the toxoplasmosis data The left panel shows fitte probabilities $\widehat{\pi}(x)$, with the fit of local linear logistic model with $h = 400$ (solid) and 0.95 pointwise confidence bands (dots). Also shown is the local linear fit with $h = 300$ (dashes). The right panel shows the loc quadratic fit with $h = 40$ and its 0.95 confidence band. Note the increased variability due to the quadratic fit, and its stronger curvature at the boundaries.

... Ex 10.32

```
> library(mgcv)
> library(SMPracticals)
> data(toxo)
> ?gam
> toxo.gam <- gam(cbind(r,m-r) ~ s(rain), family = binomial, data = toxo)
> summary(toxo.gam)

Family: binomial
Link function: logit

Formula:
cbind(r, m - r) ~ s(rain)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09015    0.08573  -1.052    0.293

Approximate significance of smooth terms:
          edf Ref.df Chi.sq p-value
s(rain) 6.515   7.57  23.05 0.00259 **

> par(mfrow=c(2,2))
> toxo.gam$sp
    s(rain)
0.008141828
> plot(gam(cbind(r,m-r) ~ s(rain),sp=toxo.gam$sp, family = binomial, data = toxo), residuals=
> plot(gam(cbind(r,m-r) ~ s(rain),sp=0.05, family = binomial, data = toxo), residuals=TRUE, p
> plot(gam(cbind(r,m-r) ~ s(rain),sp=0.5, family = binomial, data = toxo), residuals=TRUE, pc
> plot(gam(cbind(r,m-r) ~ s(rain),sp=1, family = binomial, data = toxo), residuals=TRUE, pch=
```



s(rain,6.51)
s(rain,4.68)
s(rain,2.89)
s(rain,2.46)
rain
1600
1800
2000
2200

... Ex 10.32

- `gam` uses spline smoothing terms, rather than local polynomials
- smoothing parameter replaces bandwidth *h*
- `kgplm` in library`gplm` computes kernel smooths, but for Bernoulli data

- note from output that $\phi = 1$
- `quasibinom` is a valid choice of `family`
- gives estimate of ϕ as 1.8 (with default choice of smoothing)
- smooth fit no longer significant

... Ex 10.32

- estimation of smoothing parameter using generalized cross-validation
- or generalization of AIC
- $$\text{GCV}(h) = \sum \left\{ \frac{y_j - \hat{g}(x_j)}{1 - \text{tr}(S_h)/n} \right\}^2$$
- $$\text{AIC}_c(h) = n \log \hat{\sigma}^2(h) + n \frac{1 + \text{tr}(S_h)/n}{1 - \{\text{tr}(S_h) + 2\}/n}$$
- for generalized linear models

$$\text{AIC}_c(h) = \sum d_j\{y_j; \hat{\mu}_j(h)\} + n \frac{1 + \text{tr}(S_h)/n}{1 - \{\text{tr}(S_h) + 2\}/n}$$

Flexible modelling using basis expansions §10.7.2

- $y_j = g(x_j) + \epsilon_j$
- Flexible linear modelling

$$g(x) = \Sigma_{m=1}^{M} \beta_m h_m(x)$$

- This is called a linear basis expansion, and h_m is the mth basis function
- For example if X is one-dimensional:
  $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, or
  $g(x) = \beta_0 + \beta_1 \sin(x) + \beta_2 \cos(x)$, etc.
- Simple linear regression has $h_1(x) = 1$, $h_2(x) = x$

Piecewise polynomials

- piecewise constant basis functions
  $h_1(x) = I(x < \xi_1), \quad h_2(x) = I(\xi_1 \le x < \xi_2),$
  $h_3(x) = I(\xi_2 \le x)$
- equivalent to fitting by local averaging

- piecewise linear basis functions , with constraints
  $h_1(x) = 1, \quad h_2(x) = x$
  $h_3(x) = (x - \xi_1)_+, \quad h_4(x) = (x - \xi_2)_+$
- windows defined by knots $\xi_1, \xi_2, \ldots$

- piecewise cubic basis functions
  $h_1(x) = 1, h_2(x) = x, h_3(x) = x^2, h_4(x) = x^3$

- continuity $h_5(x) = (x - \xi_1)^3_+, \quad h_6(x) = (x - \xi_2)^3_+$

- continuous function, continuous first and second derivatives

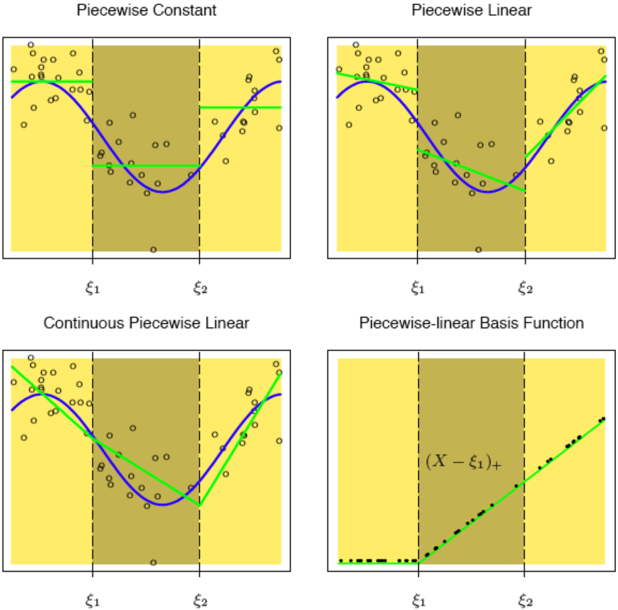


FIGURE 5.1. *The top left panel shows a piecewise constant function fit to some artificial data. The broken vertical lines indicate the positions of the two* knots *ξ_1 and ξ_2. The blue curve represents the true function, from which the data were*

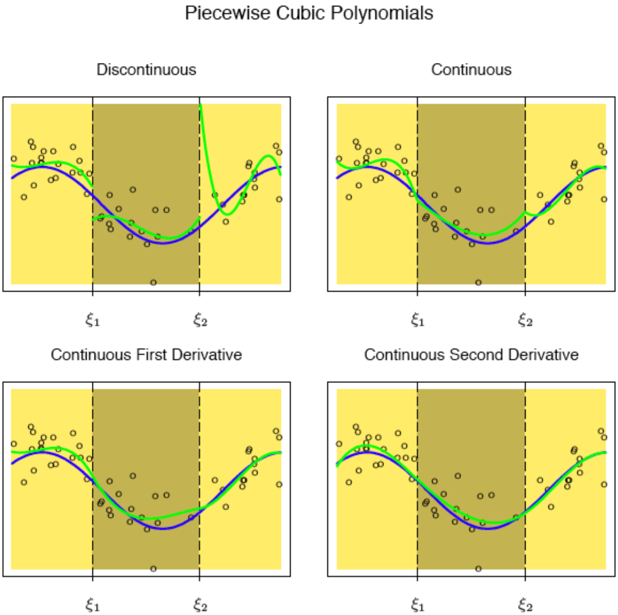
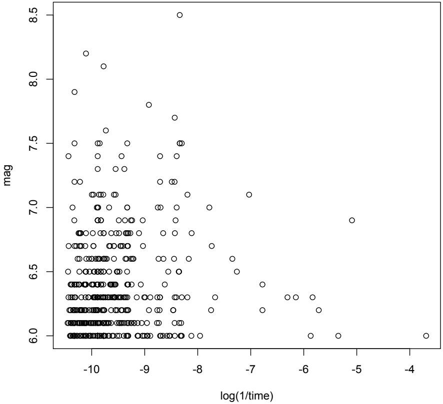


FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

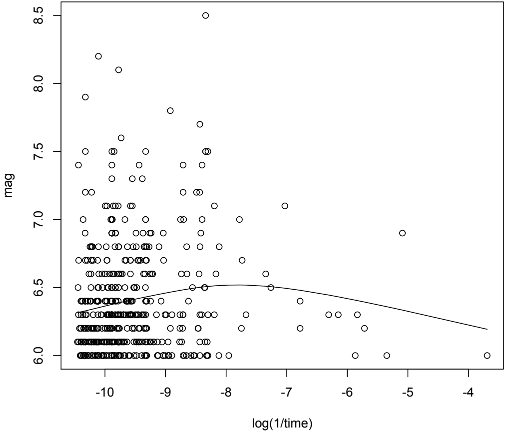
Example: earthquake data

```
> data(quake,package="SMPracticals")
> head(quake)
          time mag
1     40.08333 6.0
2    162.38889 6.9
3    210.22917 6.0
> with(quake, plot(log(1/time),mag))## using a different measure of intensity here than in Fi
```



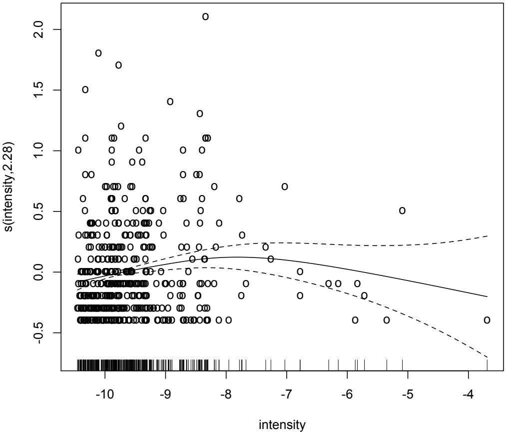
... earthquake

```
> eq.gam <- gam(mag ~ s(intensity), data = quake)
> with(quake,lines(intensity, eq.gam$fitted.values))
```



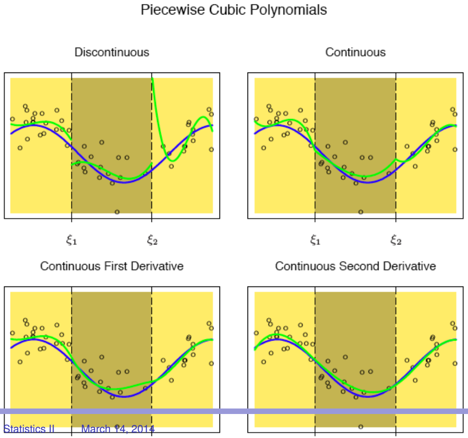
... earthquake

```
> plot(eq.gam, residual=TRUE, pch = "o")
# standard errors plotted by default
```



Cubic splines

- truncated power basis of degree 3
- need to choose number of knots K and placement of knots $\xi_1, \ldots \xi_K$ SM uses n knots
- construct features matrix using truncated power basis set
- use constructed matrix as set of predictors



... cubic splines

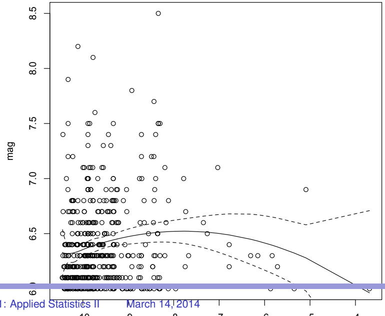
```
> with(quake, bs(log(1/time))[1:10,])
#bs(x) with no other arguments just gives a single cubic polyn
               1         2         3
 [1,] 0.0000000 0.0000000 1.0000000
 [2,] 0.1018013 0.3903714 0.4989780
 [3,] 0.1359705 0.4189773 0.4303434
 [4,] 0.1884790 0.4408886 0.3437743
 [5,] 0.2056632 0.4436068 0.3189471
...
attr(,"degree")
[1] 3
attr(,"knots")
numeric(0)
attr(,"Boundary.knots")
[1] -10.454784  -3.690961
attr(,"intercept")
[1] FALSE
attr(,"class")
[1] "bs"     "basis"  "matrix"
```

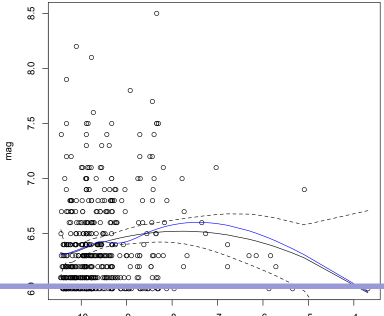
... cubic splines

```
 > with(quake,bs(log(1/time), df=5)[1:10,])
# gives a proper cubic spline basis, here with 5 df
      1          2         3         4         5
 [1,] 0 0.00000000 0.0000000 0.0000000 1.0000000
 [2,] 0 0.01110655 0.1250814 0.4247847 0.4390274
 [3,] 0 0.01846075 0.1661869 0.4486889 0.3666635
 [4,] 0 0.03370916 0.2283997 0.4600092 0.2778819
 [5,] 0 0.03989014 0.2484715 0.4585984 0.2530400
...
attr(,"degree")
[1] 3
attr(,"knots")
33.33333% 66.66667%
-9.943294 -9.520987
attr(,"Boundary.knots")
[1] -10.454784  -3.690961
```

... earthquake data

```
> quake.bs = lm(mag ~ bs(log(1/time),df=5),data = quake)
> quake.pred = predict(quake.bs, se.fit = TRUE, interval = "confidence")
> quake.pred
$fit
        fit      lwr      upr
1   5.962665 5.216283 6.709047
2   6.279641 5.979190 6.580092
3   6.323859 6.042772 6.604946
> lines(log(1/quake$time),quake.pred[[1]][,1])
> lines(log(1/quake$time),quake.pred[[1]][,2], lty=2)
> lines(log(1/quake$time),quake.pred[[1]][,3], lty=2)
> quake.lo = loess(mag ~ log(1/time), data = quake)
> quake.lopred = predict(quake.lo, se=T)
```





Smoothing splines §10.7.2

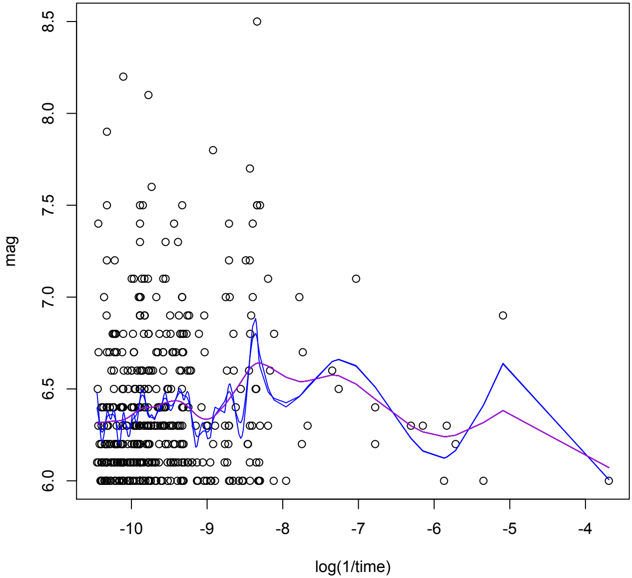
- $y_j = g(t_j) + \epsilon_j, \quad j = 1, \ldots, n$

- choose $g(\cdot)$ to solve

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

- solution is a cubic spline, with knots at each observed x_i value
- see Figure 10.18 for a non-regularized solution

- has an explicit, finite dimensional solution
- $\hat{g} = \{\hat{g}(t_1), \ldots, \hat{g}(t_n)\} = (I + \lambda K)^{-1} y$
- K is a symmetric $n \times n$ matrix of rank $n - 2$


mag
8.5
8.0
7.5
7.0
6.5
6.0
-10
-9
-8
-7
-6
-5
-4
log(1/time)

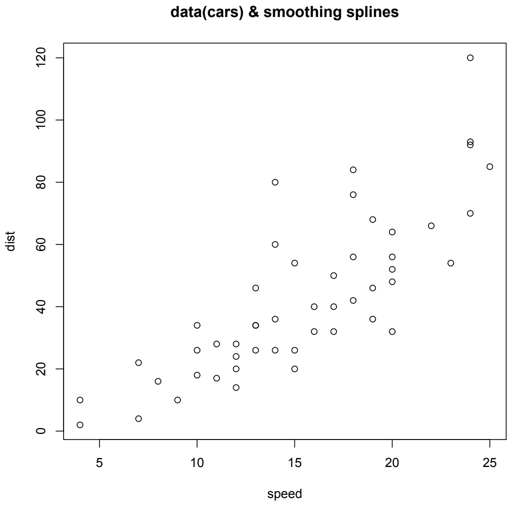
... smoothing splines

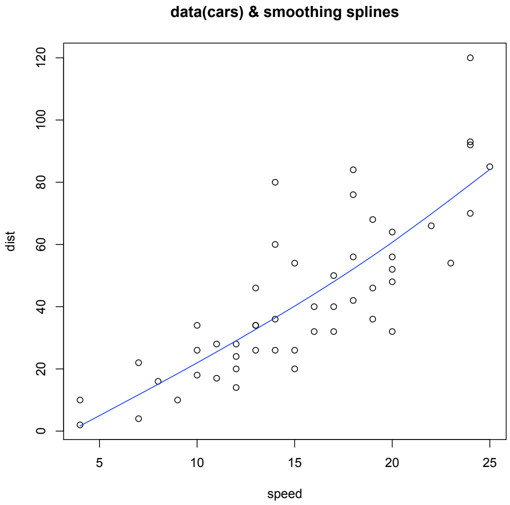
```
> quake$int = log(1/quake$time)
> quake[1:4,]
       time mag       int
1  40.08333 6.0 -3.690961
2 162.38889 6.9 -5.089994
3 210.22917 6.0 -5.348198
4 303.85417 6.2 -5.716548

> attach(quake)
> plot(int,mag)
> quake.ss2 = smooth.spline(x = int, y = mag, df = 5)
> lines(quake.ss2, col="red")
> quake.ss3
Call:
smooth.spline(x = int, y = mag, cv = TRUE)

Smoothing Parameter  spar= 1.499945  lambda= 0.0001340604 (25 iterations)
Equivalent Degrees of Freedom (Df): 11.35023
Penalized Criterion: 64.57512
PRESS: 0.1730025
> lines(quake.ss3, col="blue")
```

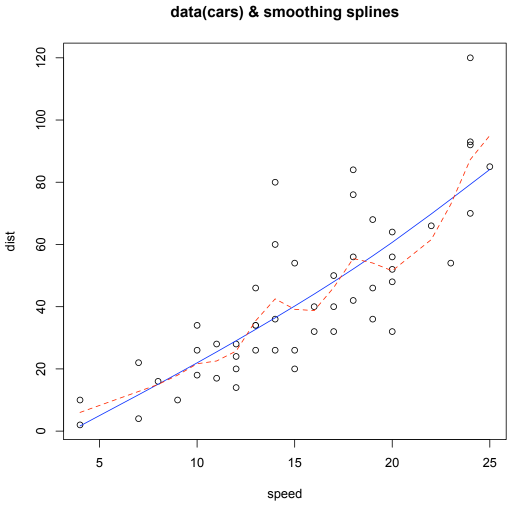
... smoothing splines

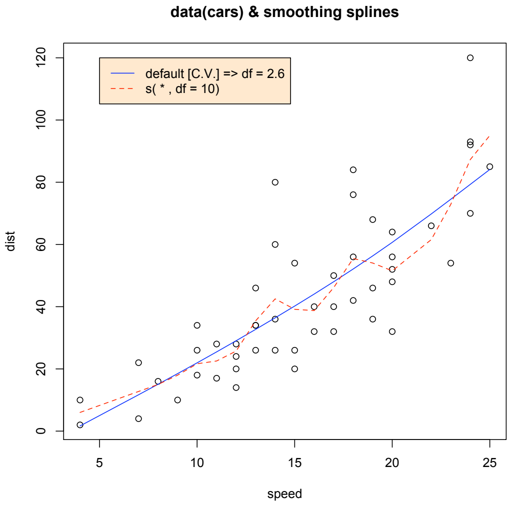
An example from the R help file for `smooth.spline`:

```
> data(cars)
> attach(cars)
> plot(speed, dist, main = "data(cars)  &  smoothing splines")
> cars.spl <- smooth.spline(speed, dist)
> (cars.spl)
Call:
smooth.spline(x = speed, y = dist)

Smoothing Parameter  spar= 0.7801305  lambda= 0.1112206 (11 iterations)
Equivalent Degrees of Freedom (Df): 2.635278
Penalized Criterion: 4337.638
GCV: 244.1044
> lines(cars.spl, col = "blue")
>      lines(smooth.spline(speed, dist, df=10), lty=2, col = "red")
>      legend(5,120,c(paste("default [C.V.] => df =",round(cars.spl$df,1)),
+                     "s( * , df = 10)"), col = c("blue","red"), lty = 1:2,
+             bg='bisque')
> detach()
```


data(cars) & smoothing splines
dist
speed
0
20
40
60
80
100
120
5
10
15
20
25


data(cars) & smoothing splines
dist
speed
0
20
40
60
80
100
120
5
10
15
20
25


data(cars) & smoothing splines
dist
speed
0
20
40
60
80
100
120
5
10
15
20
25


data(cars) & smoothing splines
default [C.V.] => df = 2.6
s(* , df = 10)
dist
speed
0
20
40
60
80
100
120
5
10
15
20
25

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
- alternative: $E(Y \mid X_1, \ldots, X_p) = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$ additive models
- binary response: $\text{logit}\{E(Y \mid X_1, \ldots, X_p)\} = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$ generalized additive models

Which smoothing method?

- basis functions: natural splines, Fourier, wavelet bases
- regularization via cubic smoothing splines
- kernel smoothers: locally constant/linear/polynomial
- adaptive bandwidth, running medians, running *M*-estimates
- Dantzig selector, elastic net, rodeo (Lafferty & Wasserman, 2008)
- Faraway (2006) Extending the Linear Model:
  - with very little noise, a small amount of local smoothing (e.g. nearest neighbours)
  - 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

link

Happy St. Patrick's Day!