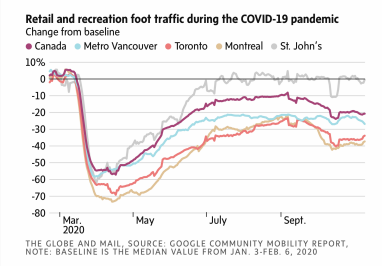
Methods of Applied Statistics I

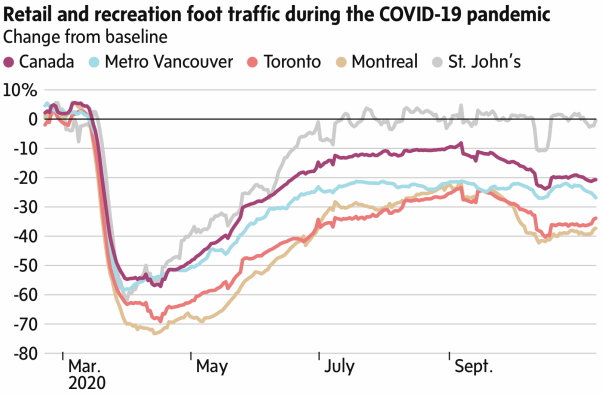
STA2101H F LEC9101

Week 10

November 19 2020



averages aren't always useful



THE GLOBE AND MAIL, SOURCE: GOOGLE COMMUNITY MOBILITY REPORT, NOTE: BASELINE IS THE MEDIAN VALUE FROM JAN. 3-FEB. 6, 2020

Today

Start Recording

1. polls again; vaccine;
2. HW3 due December 3
3. Generalized linear models summary
4. Introduction to nonparametric regression
5. Visual Inference

- November 23 15.00 – 16.00 Christine Frankline
- `https://canssiontario.utoronto.ca/?mec-events=are`
- "School level Statistics and Data Science"

Polling: November 5

Technology & Ideas

Polling Failed. It's Time to Kick the Addiction

Doubling down won't help Americans understand themselves.

By Cathy O'Neil
November 4, 2020, 2:13 PM EST

"It's no longer reasonable to assume that we can get better"

The Polls Underestimated Trump — Again. Nobody Agrees on Why.

No matter who ends up winning, the industry failed to fully account for the missteps that led it to miscalculate Donald J. Trump's support four years ago.

"there was an overestimation of Mr. Biden's support across the board – particularly with white voters and with men"

Update on polls

GUIDING FIGHT

Polling sometimes misses the mark, but without it we'd be in the dark

JEFFREY S. ROSENTHAL
The Globe and Mail (Alberta Edition)
14 Nov 2020

TProfessor of statistics at the University of Toronto whose books include Knock on Wood: Luck, Chance, and the Meaning of Everything, he U.S. presidential election gripped us all. Razor-thin margins. Four days of uncertainty about the outcome. It...
read more...

Opinion **US presidential election 2020**

Why pollsters so often seem to get it wrong

The huge win suggested for Joe Biden did not materialise — but we shouldn't expect certainty

TIM HARFORD + Add to myFT

FINANCIAL TIMES

Globe & Mail, November 14 "if Mr. Biden had instead over-performed… they probably wouldn't have complained as much "

"The pollsters are left with the challenge of forecasting how everyone will vote, based on responses from the small non-representative minority they can reach"

"polls accurately predicted the U.S. Presidential elections of 2008 and 2012, the U.S. midterm elections of 2018, the Canadian federal election of 2019…"

response rates are very low – between 1 and 5 %

"the people who do reply will be systematically different from those who do not"

"turnout in this US election has been unusually high, giving pollsters another headache"

"It's not that the polls told us nothing. It's that they could not tell us what we yearned to know"

Recap

- Generalized linear models theory
- link function, variance function, dispersion parameter ϕ, linear predictor
- derivation of maximum likelihood estimator — re-weighted LS
- examples – binomial (esoph), Poisson (cloth), Gamma (chimps)

$$\mathbb{E}(y_i) = \mu_i; \qquad g(\mu_i) = x_i^T\beta; \qquad \text{Var}(y_i) = \phi_i V(\mu_i) \qquad \phi_i = a_i\phi$$

$$\hat{\beta} = (X^TWX)^{-1}X^TWz; \quad z = X\beta + W^{-1}u; \qquad z(\beta) = X\beta + W^{-1}(\beta)u(\beta)$$

$$\text{Var}(\hat{\beta}) \doteq (X^TWX)^{-1} \qquad W \text{ is diagonal}$$

On pp. 118-119 of FELM, this iteration is carried out in `R` on the `bliss` data

Recap 2

$$\hat{\beta} = (X^T W X)^{-1} X^T W z; \quad z = X\beta + W^{-1}u; \qquad z(\beta) = X\beta + W^{-1}(\beta)u(\beta)$$

$$\text{Var}(\hat{\beta}) \doteq (X^T W X)^{-1}$$

W is diagonal

$$W_{ii} =$$

$$u_i =$$

Note $\hat{\beta}$ is free of ϕ because of W and W^{-1}, but $\text{Var}(\hat{\beta})$ depends on ϕ

Warning: in FELM W is defined slightly differently (no ϕ), so he has $\text{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$

Recap 2

$$\hat{\beta} = (X^T W X)^{-1} X^T W z; \quad z = X\beta + W^{-1}u; \qquad z(\beta) = X\beta + W^{-1}(\beta)u(\beta)$$
$$\mathrm{Var}(\hat{\beta}) \doteq (X^T W X)^{-1} \qquad W \text{ is diagonal}$$

$$W_{ii} = \frac{1}{\phi a_i \{g'(\mu_i)\}^2 V(\mu_i)}$$

$$u_i = \frac{y_i - \mu_i}{\phi a_i g'(\mu_i) V(\mu_i)}$$

Note $\hat{\beta}$ is free of ϕ because of W and W^{-1}, but $\mathrm{Var}(\hat{\beta})$ depends on ϕ

Warning: in FELM W is defined slightly differently (no ϕ), so he has $\mathrm{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$
Further, the w_i on p.117 is not the same as the w_o on p. 188; SM uses a_i instead which would have been better for FELM

The last slide about GLM theory

- choose a model, often based on type of response — or on mean/variance relationship
- fit a model, using maximum likelihood estimation — convergence (almost) guaranteed
- inference for individual coefficients $\hat{\beta}_j$ from `summary`
- inference for groups of coefficients by analysis of deviance

- estimation of ϕ based on Pearson's Chi-square

another typo in FELM p.121: cross out $= \text{var}(\hat{\mu})$

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

- analysis of deviance: see p. 121 (near bottom) — likelihood ratio tests
- diagnostics: same as for `lm` — FELM p.124; SM p.477
  - residuals: deviance or Pearson; can be standardized — FELM likes 1/2 normal plots
  - influential observations: uses hat matrix — `SMPracticals` has very good GLM diagnostics

`glm.diag, plot.glm.diag`

Really the last slide about GLM theory

- special to `glm`
- two models, Poisson and Binomial, have no ϕ parameter
- this has two consequences
- the residual deviance can be used as a test of fit of the model
- two pseudo-models are available called `quasibinomial, quasipoisson`
- quasi-binomial: $\text{var}(y_i) = \phi p_i(1 - p_i)$
- quasi-Poisson: $\text{var}(y_i) = \phi \mu_i$
- quasi- is a quick way to fit proportion or count responses, but allow the variance to be bigger (or rarely, smaller) than it would be under the binomial or Poisson model
- caveat – none of this works for binary data, only binomial

$n_i \geq 5$, approx

- model $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \quad x_i$ scalar
- mean function $f(\cdot)$ assumed to be "smooth"
- introduce a kernel function $K(u)$ and define a set of weights

$$w_i = \frac{1}{\lambda} K\left(\frac{x_i - x_0}{\lambda}\right)$$

- estimate of $f(x)$, at $x = x_0$:

$$\hat{f}_\lambda(x_0) = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$

- Nadaraya-Watson estimator – local averaging

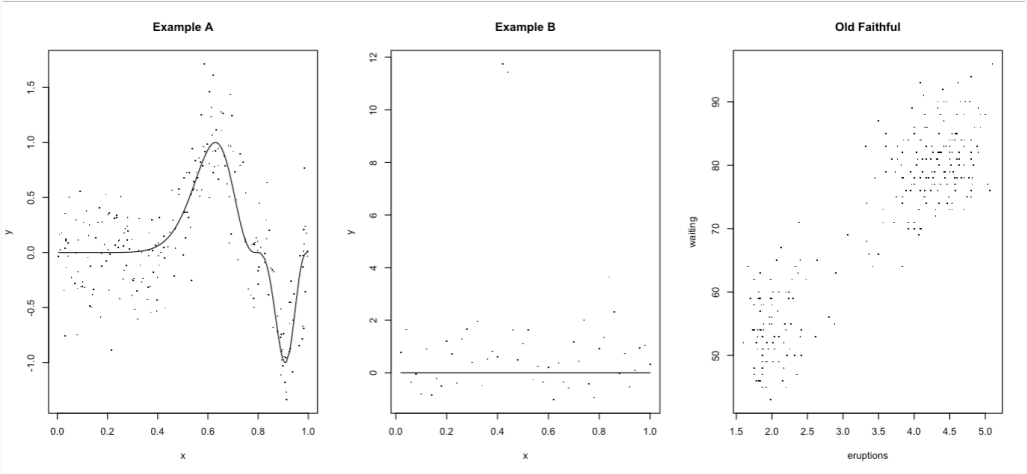
local polynomial of degree 0

Kernel smoothers

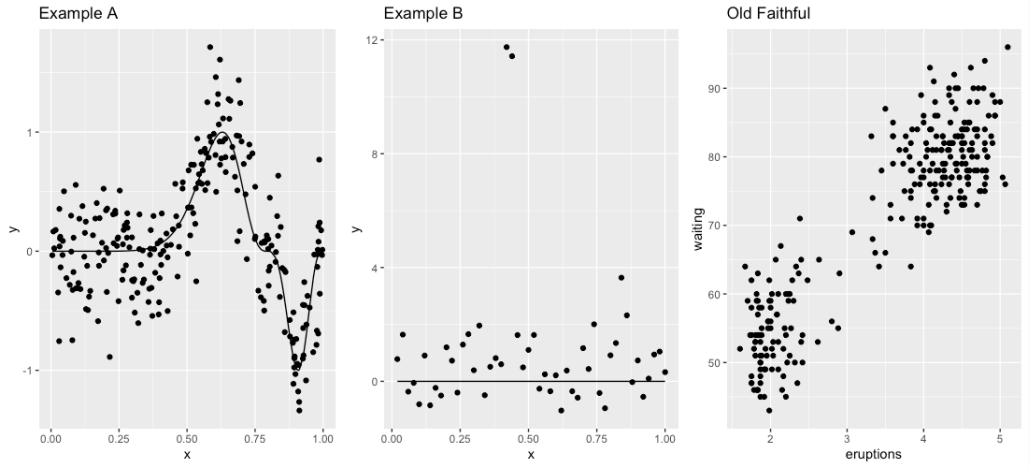
- choice of bandwidth, λ controls smoothness of function
- larger bandwidth $=$ more smoothing
- kernel estimators are biased
- making the estimate smoother increases bias, decreases variance

- choice of kernel function, $K(\cdot)$, controls smoothness and "local-ness"
- Faraway recommends Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2), |x| \leq 1$
- `ksmooth(base)` offers only uniform (box) or normal
- `bkde(KernSmooth)` offers `normal, box, epanech, biweight, triweight`
- `biweight`: $K(x) = (1 - |x|^2)^3, |x| \leq 1$

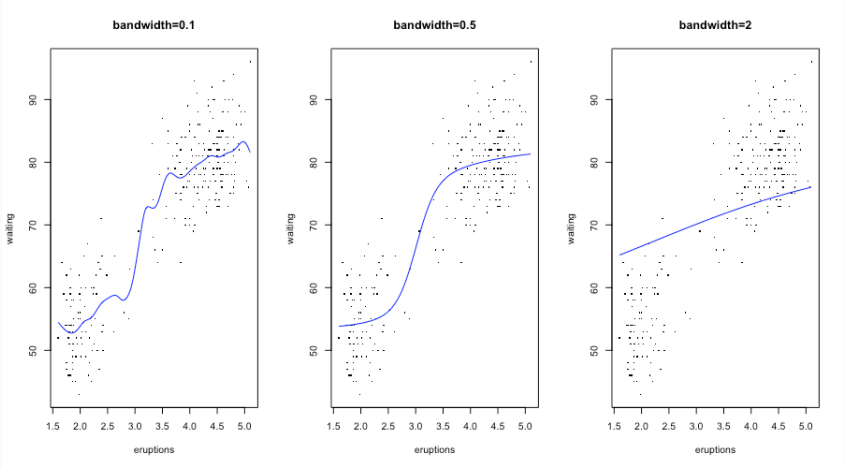
Examples



```
exb <- data.frame(exb)

plota <- ggplot(exa) + geom_point(aes(x,y)) +
geom_line(aes(x,m))+ ggtitle("Example A")

plotb <- ggplot(exb) + geom_point(aes(x,y)) +
geom_line(aes(x,m))+ ggtitle("Example B")

plotc <- ggplot(faithful) + geom_point(aes(eruptions,waiting)) +
 ggtitle("Old Faithful")

grid.arrange(plota, plotb, plotc, nrow=1) #in gridExtra library
```


Example A
y
x
1
0
-1
0.00
0.25
0.50
0.75
1.00
Example B
y
x
12
8
4
0
0.00
0.25
0.50
0.75
1.00
Old Faithful
waiting
eruptions
90
80
70
60
50
2
3
4
5

```
with(faithful, plot(eruptions, waiting, cex=2, main = "bandwidth=0.1", pch="."))
lines(locpoly(faithful$eruptions,faithful$waiting,drv=0L,
  degree=0,bandwidth=.1), col = "blue")

with(faithful, plot(eruptions, waiting, cex=2, main = "bandwidth=0.5", pch="."))
lines(locpoly(faithful$eruptions,faithful$waiting,drv=0L,
  degree=0, bandwidth=.5), col = "blue")

with(faithful, plot(eruptions, waiting, cex=2, main = "bandwidth=2", pch="."))
lines(locpoly(faithful$eruptions,faithful$waiting,drv=0L,
  degree=0, bandwidth=2), col = "blue")
```



These are smoother than the plots in FELM using `base::ksmooth`

Bias and MSE

- Nadaraya-Watson: $\hat{f}_\lambda(x) = \Sigma w_i y_i / \Sigma w_i$; $\quad w_i = \frac{1}{\lambda} K(\frac{x_i - x_0}{\lambda})$

- $\hat{f}_\lambda(x)$ is biased

$$\mathrm{E}\{\hat{f}_\lambda(x)\} \doteq \frac{1}{2}\lambda^2 f''(x)$$

$$\mathrm{var}\{\hat{f}_\lambda(x)\} \doteq \frac{\sigma^2}{n\lambda f_\lambda(x)} \int K^2(u)du$$

- could choose λ to minimize MSE $=$ bias^2 $+$ var, at x
- could choose λ to minimize integrated MSE

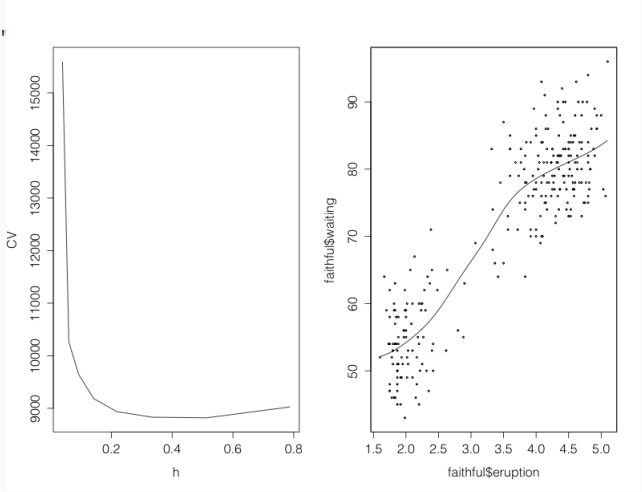
- more usual to use cross-validation

SM 10.7.1 (no n); FELM 11.1

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}_{-i}(x_i)\}^2$$

Cross-validation

```
library(sm)
hm <- hcv(faithful$eruptions,
   faithful$waiting, display = "lines"
sm.regression(faithful$eruptions,
   faithful$waiting, h = hm,
   xlab = "eruptions",
   ylab = "waiting")
```



- above uses local averaging based on kernel function
- better estimates can be obtained using local regression at point x

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^k \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x_0) & \cdots & (x_n - x_0)^k \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

-

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

-

$$\hat{f}_\lambda(x_0) = \hat{\beta}_0$$

- usually evaluate the function at sample points: $\hat{f}_\lambda(x_i), i = 1, \ldots, n$

- odd-order polynomials work better than even; usually local linear fits are used
- kernel function is often a Gaussian density, or the tricube kernel

$$K(u) = (1 - |u|^3)^3, \quad |u| \leq 1$$

- as with N-W (local averaging) estimators, choice of bandwidth controls smoothness
- **loess** is the most widely used, and is the default in `ggplot2`
- fits a local linear regression, but not by least squares
- uses a robust version of least squares that downweights outliers
- the result is that the bandwidth can change with x

- $\hat{\beta} = (X^T W X)^{-1} X^T W y \qquad W = \text{diag}(w_1, \ldots, w_n)$
- $\hat{f}_\lambda(x_0) = \hat{\beta}_0 = \sum_{i=1}^n S(x_0; x_i, \lambda) y_i$
- $S(x_0; x_1, \lambda), \ldots, S(x_0; x_n, \lambda)$ first row of "hat" matrix
- this makes it relatively easy to analyse the behaviour of local polynomial smoothers

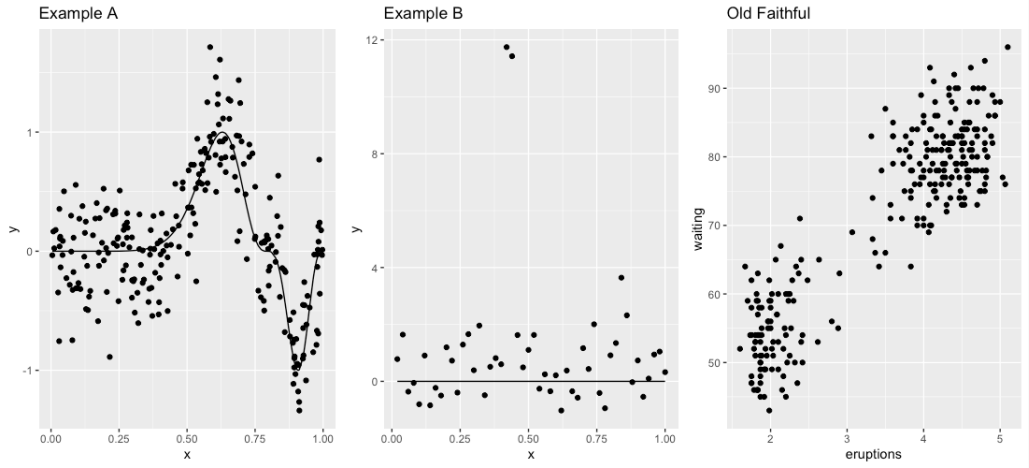
SM §10.7

- and to simplify the expression for the cross-validation criterion $CV(\lambda)$
- fitting at each sample value gives

$$\hat{f}_\lambda(x_i) = \sum_{j=1}^n S(x_i; x_j, \lambda) y_j$$

- $$CV(\lambda) = \sum_{i=1}^{n} \{y_i - \hat{f}_{-i}(x_i)\}^2$$
- for local polynomials
$$CV(\lambda) = \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}(\lambda)} \right\}^2$$
- even simpler
$$GCV(\lambda) = \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathrm{tr}(S_\lambda)/n} \right\}^2$$
- $$\hat{f}_\lambda(x_i) = \sum_{j=1}^{n} S(x_i; x_j, \lambda) y_j$$

Examples


Example A
y
x
1
0
-1
0.00
0.25
0.50
0.75
1.00
Example B
y
x
12
8
4
0
0.00
0.25
0.50
0.75
1.00
Old Faithful
waiting
eruptions
90
80
70
60
50
2
3
4
5

Examples

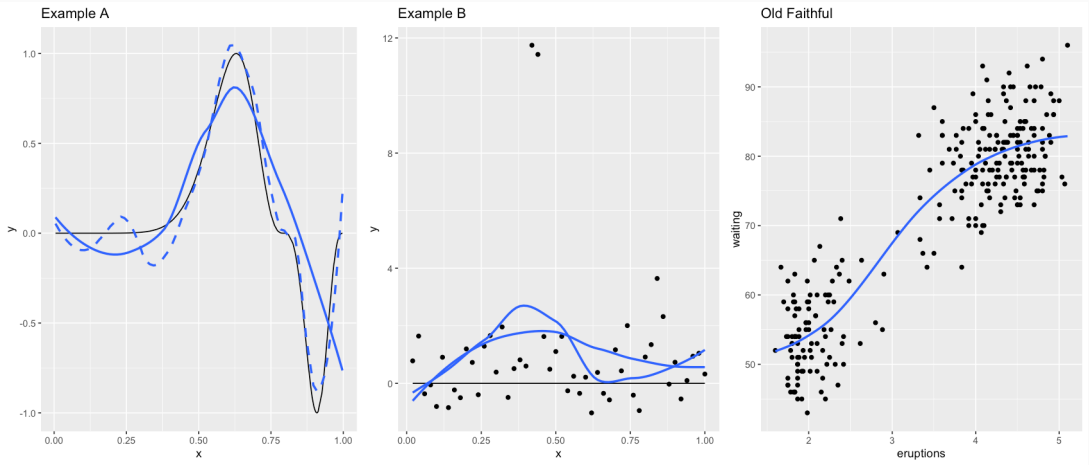
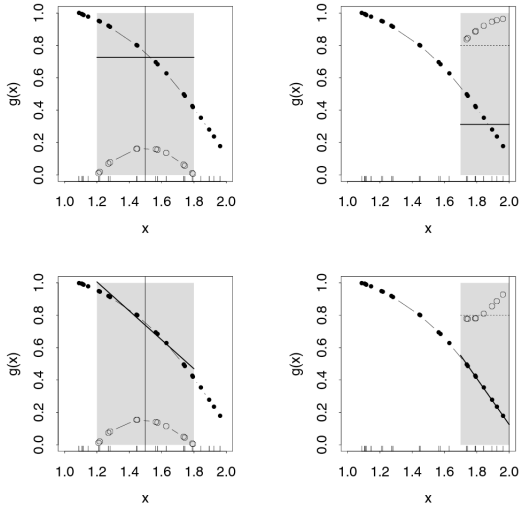
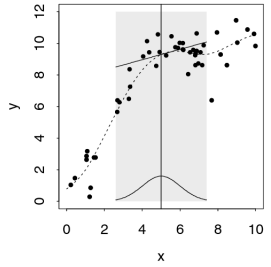

Example A
1.0
0.5
0.0
-0.5
-1.0
y
0.00
0.25
0.50
0.75
1.00
x
Example B
12
8
4
0
y
0.00
0.25
0.50
0.75
1.00
x
Old Faithful
90
80
70
60
50
waiting
2
3
4
5
eruptions

Figure 10.15 Local polynomial fitting by least squares. In each panel the function $g(x)$ is shown by a line joining the solid blobs (x_j, y_j), shown without error for clarity, and the target value x_0 at which g is to be estimated is given by the vertical line; $x_0 = 1.5$ for the left panels and $x_0 = 2$ for the right panels. Only observations falling inside the shaded region contribute to the fit, and the effective kernel is shown by the circles; in the right panels the effective kernel has been shifted upwards by 0.8. The heavy solid lines show the local polynomial fits, which are constant in the upper panels and linear in the lower panels. The local constant fit is more biased than the local linear fit, especially at the edge $x_0 = 2$.





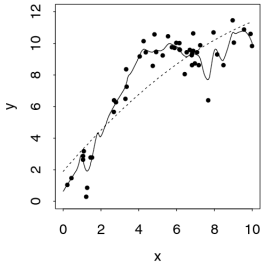


Figure 10.14
Construction of a local linear smoother. Left panel: observations in the shaded part of the panel are weighted using the kernel shown at the foot, with $h = 0.8$, and the solid straight line is fitted by weighted least squares. The local estimate is the fitted value when $x = x_0$, shown by the vertical line. Two hundred local estimates formed using equi-spaced x_0 were interpolated to give the dotted line, which is the estimate of $g(x)$. Right panel: local linear smoothers with $h = 0.2$ (solid) and $h = 5$ (dots).

Recall that a kernel function $w(u)$ is a unimodal density function symmetric about $u = 0$ and with unit variance. One choice of w is the standard normal density. Another is a rescaled form of the *tricube* function

$$w(u) = \begin{cases} (1 - |u|^3)^3, & |u| \le 1, \\ 0, & \text{otherwise}, \end{cases} \qquad (10.37)$$

and there are many others.

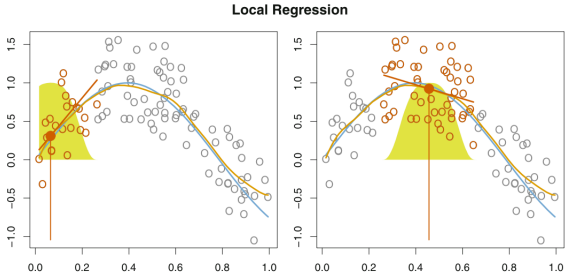
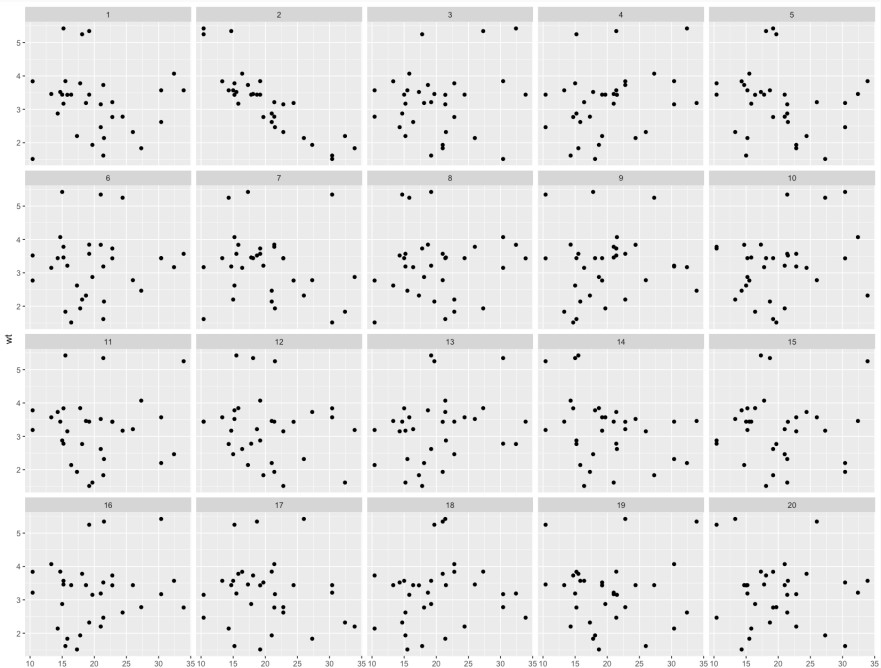
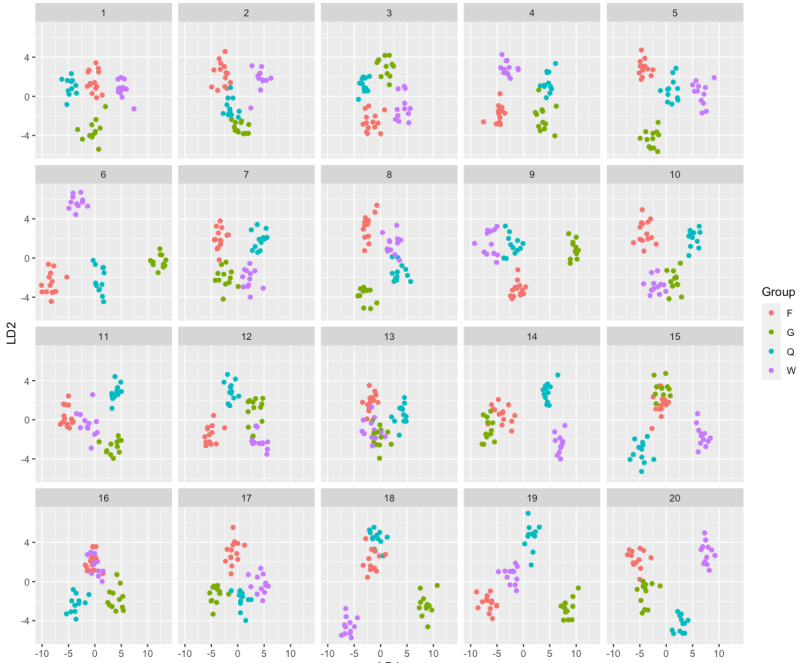


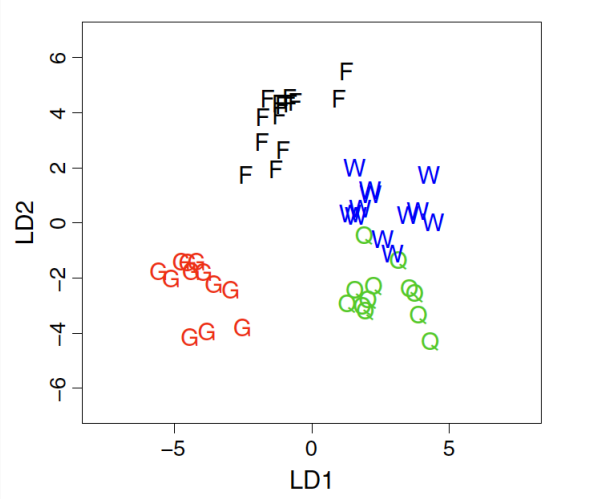
FIGURE 7.9. *Local regression illustrated on some simulated data, where the blue curve represents $f(x)$ from which the data were generated, and the light orange curve corresponds to the local regression estimate $\hat{f}(x)$. The orange colored points are local to the target point x_0, represented by the orange vertical line. The yellow bell-shape superimposed on the plot indicates weights assigned to each point, decreasing to zero with distance from the target point. The fit $\hat{f}(x_0)$ at x_0 is obtained by fitting a weighted linear regression (orange line segment), and using the fitted value at x_0 (orange solid dot) as the estimate $\hat{f}(x_0)$.*

Lineups


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
wt
5
4
3
2
10
15
20
25
30
35


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
LD2
4
0
-4
-10
-5
0
5
10
Group
F
G
Q
W

Chowdhury et al. 2015, *Comput. Statist.*



Class ended at this point

the following slides will be for next week

How do we choose a model?

CD Ch. 6.5

- in many fields of study the models used as a basis for interpretation do not have a special subject-matter base
- rather represent broad patterns of haphazard variation quite widely seen
- this is typically combined with a specification of the systematic part of the variation
- which is often the primary focus
- modelling then often reduces to a choice of distributional form
- and of the independence structure of the random components

- functional form of the probability distribution sometimes critical, for example where an implicit assumption is involved of a relationship between variance and mean: geometric, Poisson, binomial
- the simple situations that give rise to binomial, Poisson, geometric, exponential, normal and log normal are some guide to empirical model choice in more complex situations

- In some specific contexts there is a tradition establishing the form of model
- illustration: financial time series – $Y(t) = \log\{P(t)/P(t-1)\}$ has a long-tailed distribution, small serial correlation, large serial correlation in $Y^2(t)$ stock price

- often have a long tail of large values; exponential distribution is a natural staring point
- extensions may be needed, including Weibull, gamma or log-normal

- often helpful to develop random and systematic parts of the model separately
- models should obey natural or known constraints, even if these lie outside the range of the data
- example $P(Y = 1) = \alpha + \beta x \implies \log \frac{P(Y=1)}{P(Y=0)} = \alpha' + \beta' x$

- however, β measures the change in probability per unit change in x β' does not
- when relationship between y and several variables $x_1, \ldots x_p$ is of interest
  - unlikely that the system is wholly linear
  - impractical to study nonlinear systems of unknown form
  - therefore reasonable to begin with a linear model
  - and seek isolated nonlinearities

... empirical models

CD Ch. 6.5

- often helpful to develop random and systematic parts of the model separately
- naive approach: one random variable per study individual
- values for different individuals independent

- more realistic: possibility of structure in the random variation
- dependence in time or space, or a hierarchical structure corresponding to levels of aggregation
- ignoring these complications may give misleading assessments of precision, or bias the conclusions

- example: standard error of mean $\sigma/\sqrt{n}$
- but, under mutual correlation, becomes $(\sigma/\sqrt{n})(1+\Sigma\rho_{ij})^{1/2}$
- if each observation correlated with k others, at same level, $(\sigma/\sqrt{n})(1+k\rho)^{1/2}$

0.1	0.2	0.4	0.8
1.14	1.26	1.48	1.84
1.18	1.34	1.61	2.05
1.22	1.41	1.73	2.24
1.26	1.48	1.84	2.41
1.30	1.55	1.95	2.57
1.34	1.61	2.05	2.72

$k = 3:8$

- important to be explicit about the unit of analysis
- has a bearing on independence assumptions involved in model formulation
- example: if all patients in the same clinic receive the same treatment
- then the clinic is the unit of analysis

- in some contexts there may be a clear hierarchy
- assessment of precision comes primarily from comparisons between units
- modelling of variation within units is necessary only if of intrinsic interest

- when relatively complex responses are collected on each individual, the simplest way of condensing these is through a number of summary descriptive measures
- in other situations it may be necessary to represent explicitly the different hierarchies of variation

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, ...

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

NMMAPS: National Morbidity, Mortality and Air Pollution Study

... 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). *Biometrika* **95** 635–51