Regression Methods

Anthony Davison

http://stat.epfl.ch

Dictionary

- **Regression**: (statistics) a measure of the relation between the mean value of
  - one variable (e.g., output), denoted y (the **response variable**) and
  - corresponding values of other variables (e.g., time and cost), denoted x (**explanatory variables**).
- The explanatory variables are also called **covariates** or **features** (ML).
- We avoid the terms **dependent variable** (Y) and **independent variable** (x) used in older books.

- Questions we try and answer:
  - (**description/explanation**) how does y depend on x? How much of the variation of y is due to x? Do I need all of x to explain the variation in y?
  - (**prediction**) what will y be if $x = x_+$?
  - (**causation**) if I change x, what will happen to y?
- The causation question presupposes that we can change (some of) x, which is not always true.

Linear model

- Simplest explanation of y in terms of x is **linear model**:

$$y = g(x) = x_1\beta_1 + \cdots + x_p\beta_p = x^{\mathrm{T}}\beta,$$

  where

$$y \in \mathbb{R}, \quad x^{\mathrm{T}} = (x_1, \ldots, x_p) \in \mathbb{R}^p, \quad \beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p.$$

- The data consist of n **instances/examples/cases** (x_j, y_j) for $j = 1, \ldots, n$, so

$$y_{n\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta_{p\times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

  and we write

$$y = X\beta.$$

- **Key point**: linearity refers to linearity in β, not in terms of elements of X, which might be polynomials, or basis functions, or ...
- Sometimes we can transform to a linear model. For example, the multiplicative expression $y = \gamma x_1^{\beta_1} x_2^{\beta_2}$ becomes

$$\log y = \log \gamma + \beta_1 \log x_1 + \beta_2 \log x_2.$$

Notation

- ☐ Vectors are column vectors
- ☐ We sometimes write $X_{n\times p}$ to give the dimensions of a matrix or vector
- ☐ a^{T} (row vector) is the transpose of a (column vector)
- ☐ $j \in \{1,\ldots,n\}$ (or sometimes i) indexes the rows of y (cases/examples)
- ☐ x_j^{T} is the jth row of X
- ☐ $r, s, t, \ldots \in \{1,\ldots,p\}$ indexes the columns of X (covariates/features)
- ☐ Roman letters (y, X, z, $\ldots$) denote observed quantities, and may be the realisations of random variables
- ☐ Greek letters ($\beta, \gamma, \theta, \sigma, \ldots$) denote unknown (often vector) parameters of models
- ☐ $\widehat{\beta}$ denotes an estimate of β
- ☐ α denotes the level of significance tests and confidence intervals
- ☐ If Q is scalar (or a row vector) and β is a vector, then $\partial Q/\partial\beta$ denotes the vector (or matrix) the same shape as β with elements $\partial Q/\partial\beta_r$.
- ☐ If Q is scalar and β, γ are vectors, then $\partial^2 Q/\partial\beta\partial\gamma^{\mathrm{T}}$ denotes the matrix with (r, s) element $\partial^2 Q/\partial\beta_r\partial\gamma_s$.
- ☐ $Y \perp\!\!\!\perp Z$ means that the random variables Y and Z are independent

Least squares fit

- ☐ Assume that

$$y = X\beta$$

and find the 'best fit' by choosing β to minimise the (squared) Euclidean distance between y and $X\beta$, i.e., the sum of squares

$$\|y - X\beta\|^2 = (y - X\beta)^{\mathrm{T}}(y - X\beta) = \sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\beta)^2.$$

- ☐ In vector space terms, $y \in \mathbb{R}^n$ and $X\beta \in \mathrm{span}(X) \subset \mathbb{R}^n$.
- ☐ The 'best fit' vector $\widehat{y}$ is the vector in $\mathrm{span}(X)$ closest to y; Pythagoras' theorem (sketch) gives $\widehat{y} \perp (y - \widehat{y})$ (but see below).
- ☐ Below we call $\widehat{y} \in \mathbb{R}^n$ the **fitted value(s)** and $e = y - \widehat{y} \in \mathbb{R}^n$ the **residual (vector)**.

Lemma 1 *Without loss of generality X has rank p. If $n \geq p$ then $\widehat{y} = X\widehat{\beta} = Hy$, where*

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \quad H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}.$$

The 'hat matrix' H has rank p, is symmetric and idempotent, and satisfies $HX = X$: it gives the orthogonal projection of $\mathbb{R}^n$ onto $\mathrm{span}(X)$.

Note to Lemma 1

- ☐ If $X_{n\times p}$ does not have full column rank, then there exists a linearly independent subset of columns, X', say, such that $X\beta = X'\gamma$ for every β, and we would then minimise $\|y - X'\gamma\|^2$. Hence there is no loss of (mathematical) generality in supposing that X has full column rank.
- ☐ The sum of squares

$$Q = (y - X\beta)^{\mathrm{T}}(y - X\beta) = y^{\mathrm{T}}y - \beta^{\mathrm{T}}X^{\mathrm{T}}y - y^{\mathrm{T}}X\beta + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta = y^{\mathrm{T}}y - 2y^{\mathrm{T}}X\beta + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta$$

has first and second derivatives (respectively a $p \times 1$ vector and $p \times p$ matrix)

$$\frac{\partial Q}{\partial \beta} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\beta, \quad \frac{\partial^2 Q}{\partial\beta\partial\beta^{\mathrm{T}}} = 2X^{\mathrm{T}}X$$

with respect to β. Setting $\partial Q/\partial\beta = 0$ implies that $(X^{\mathrm{T}}X)\beta = X^{\mathrm{T}}y$, and as $X^{\mathrm{T}}X$ has rank p it is invertible, so we can write

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \quad \widehat{y} = X\widehat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Hy,$$

say. The matrix $X^{\mathrm{T}}X$ is positive definite, so $(y - X\beta)^{\mathrm{T}}(y - X\beta)$ is minimised at $\widehat{\beta}$.
- ☐ The $n \times n$ 'hat matrix' H (which 'puts a hat' on y) satisfies $H^{\mathrm{T}} = H$, $H^2 = H$, so it is symmetric and idempotent, i.e., its eigenvalues equal 0 or 1, and their multiplicities must be $n - p$ and p, as its rank is p. H is the matrix that projects $\mathbb{R}^n$ orthogonally onto $\mathrm{span}(X)$.
- ☐ The inner product between $\widehat{y}$ and $y - \widehat{y}$ equals zero, because $\widehat{y} = Hy$, $y - \widehat{y} = (I - H)y$, and $\widehat{y}^{\mathrm{T}}(y - \widehat{y}) = y^{\mathrm{T}}H^{\mathrm{T}}(I - H)y = y^{\mathrm{T}}(H - H)y = 0$.
- ☐ Clearly $HX = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X = X$, and $H(X\beta) = X\beta$ for any $\beta \in \mathbb{R}^p$.

Analysis of variance

- ☐ Let X' be a subset of the columns of X, and let $\widehat{y}'$ be the corresponding best fit. Then

$$y = \widehat{y}' + (\widehat{y} - \widehat{y}') + (y - \widehat{y}) = H'y + (H - H')y + (I - H)y$$

(in an obvious notation) where (Pythagoras again)

$$\widehat{y}' \quad \perp \quad \widehat{y} - \widehat{y}' \quad \perp \quad y - \widehat{y},$$

which implies that

$$H'(H - H') = H'(I - H) = H'(I - H') = H(I - H) = 0, \quad H'H = H'.$$

- ☐ This gives the **analysis of variance (ANOVA)**

$$\|y\|^2 = \|\widehat{y}'\|^2 + \|\widehat{y} - \widehat{y}'\|^2 + \|y - \widehat{y}\|^2$$

which decomposes ('analyses') the total variation $\|y\|^2$ of y into
  - the contribution $\|\widehat{y}'\|^2$ due to the columns of X',
  - the contribution $\|\widehat{y} - \widehat{y}'\|^2$ due to the columns of X additional to the columns of X',
  - the **residual sum of squares** $\|y - \widehat{y}\|^2$ left once the columns of X have been fitted.
- ☐ Clearly this generalises to vectors of fitted values from $X', X'', X''', \ldots$

Coefficient of determination

- **Coefficient of determination** R^2 measures reduction in variance of y as

$$R^2 = \frac{\|\widehat{y} - \overline{y}1_n\|^2}{\|y - \overline{y}1_n\|^2} = \frac{\{(I - H_1)\widehat{y})\}^{\mathrm{T}}(I - H_1)\widehat{y}}{\{(I - H_1)y)\}^{\mathrm{T}}(I - H_1)y} = \frac{y^{\mathrm{T}}(H - H_1)y}{y^{\mathrm{T}}(I - H_1)y},$$

where H_1 and H are the hat matrices for regression on 1_n and X, and $1_n \in \mathrm{span}(X)$.
- $R^2 \in [0, 1]$ is the squared empirical correlation between y and $\widehat{y}$, so $R^2 \approx 1$ implies that most of the variation in y is explained by $\widehat{y}$.
- There is a geometric interpretation, as the terms on the right of

$$(I_n - H_1)y = (I_n - H)y + (H - H_1)y$$

are orthogonal.
- Adding columns to X must increase R^2, so often use **adjusted** R^2,

$$R_a^2 = R^2 + (1 - R^2)\frac{n - 1}{n - p}.$$

- If $1_n \notin \mathrm{span}(X)$, use

$$R_0^2 = \frac{\widehat{y}^{\mathrm{T}}\widehat{y}}{y^{\mathrm{T}}y}, \quad R_{0,a}^2 = R_0^2 + (1 - R_0^2)\frac{n}{n - p}.$$

Reminders slide 9

Moment-generating function

Definition 2 *The* **moment-generating function (MGF)** *of a random vector* $Y_{n\times 1}$ *is*

$$M_Y(t) = \mathrm{E}(e^{t^{\mathrm{T}}Y}) = \mathrm{E}(e^{\sum_{j=1}^n t_j Y_j}), \quad t \in \mathcal{T} = \{t \in \mathbb{R}^n : M_Y(t) < \infty\},$$

and the **cumulant-generating function** *of* Y *is* $K_Y(t) = \log M_Y(t)$, $t \in \mathcal{T}$.

Then
- $0 \in \mathcal{T}$, so $M_Y(0) = 1$ and $K_Y(0) = 0$;
- if $\mathcal{T}$ contains an open set, then

$$\mu = \mathrm{E}(Y) = K_Y'(0) = \left.\frac{\partial K_Y(t)}{\partial t}\right|_{t=0}, \quad \Omega = \mathrm{var}(Y) = \left.\frac{\partial^2 K_Y(t)}{\partial t \partial t^{\mathrm{T}}}\right|_{t=0};$$

- if $\mathcal{A}, \mathcal{B}$ are disjoint subsets of $\{1, \ldots, n\}$ and $Y_{\mathcal{A}}$ denotes the sub-vector of Y containing $\{Y_j : j \in \mathcal{A}\}$, etc., then $Y_{\mathcal{A}} \perp\!\!\!\perp Y_{\mathcal{B}}$ if and only if

$$M_Y(t) = \mathrm{E}(e^{t_{\mathcal{A}}^{\mathrm{T}}Y_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}Y_{\mathcal{B}}}) = M_{Y_{\mathcal{A}}}(t_{\mathcal{A}})M_{Y_{\mathcal{B}}}(t_{\mathcal{B}}), \quad t \in \mathcal{T};$$

- the MGF of $Y_{\mathcal{A}}$ equals $M_Y(t)$ evaluated with $t_{\mathcal{B}} = 0$;
- if we recognise an MGF, then we know the probability distribution that gave it.

Multivariate normal distribution

Definition 3 *The random vector* $Y = (Y_1, \dots, Y_n)^{\mathrm{T}}$ *has a* **multivariate normal distribution**, $Y \sim \mathcal{N}_n(\mu, \Omega)$, *if there exist a vector* $\mu = (\mu_1, \dots, \mu_n)^{\mathrm{T}} \in \mathbb{R}^n$ *and a symmetric matrix* $\Omega \in \mathbb{R}^{n \times n}$ *with elements* ω_{jk} *such that*

$$u^{\mathrm{T}}Y \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u), \quad u \in \mathbb{R}^n.$$

Theorem 4 *If* $Y \sim \mathcal{N}_n(\mu, \Omega)$, *then*
(a)

$$\mathrm{E}(Y) = \mu, \quad \mathrm{var}(Y) = \Omega;$$

(b) the MGF of Y *is* $M_Y(t) = \exp(t^{\mathrm{T}}\mu + \frac{1}{2}t^{\mathrm{T}}\Omega t)$, $t \in \mathbb{R}^n$;
(c) if $\mathcal{A}, \mathcal{B}$ *are disjoint subsets of* $\{1, \dots, n\}$, *then*

$$Y_{\mathcal{A}} \perp\!\!\!\perp Y_{\mathcal{B}} \quad \Leftrightarrow \quad \Omega_{\mathcal{A},\mathcal{B}} = 0,$$

where $\Omega_{\mathcal{A},\mathcal{B}}$ *contains* ω_{ij} *for* $i \in \mathcal{A}$, $j \in \mathcal{B}$;
(d) if $Y_1, \dots, Y_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, *then* $Y = (Y_1, \dots, Y_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$; *and*
(e) linear functions of Y *are normally distributed, i.e.,*

$$a_{m \times 1} + B_{m \times n}Y \sim \mathcal{N}_m(a + B\mu, B\Omega B^{\mathrm{T}}).$$

Note to Theorem 4

(a) Let e_j denote the n-vector with 1 in the $j\mathrm{th}$ place and zeros everywhere else. Then $Y_j = e_j^{\mathrm{T}} Y \sim N(\mu_j, \omega_{jj})$, giving the mean and variance of Y_j.
Now $\mathrm{var}(Y_j + Y_k) = \mathrm{var}(Y_j) + \mathrm{var}(Y_k) + 2\mathrm{cov}(Y_j, Y_k)$, and

$$Y_j + Y_k = (e_j + e_k)^{\mathrm{T}} Y \sim \mathcal{N}(\mu_j + \mu_k, \omega_{jj} + \omega_{kk} + 2\omega_{jk}),$$

which implies that $\mathrm{cov}(Y_j, Y_k) = \omega_{jk} = \omega_{kj}$.

(b) Since $u^{\mathrm{T}} Y \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u)$, its MGF is $M_{u^{\mathrm{T}}Y}(t) = \mathrm{E}(e^{tu^{\mathrm{T}}Y}) = \exp(tu^{\mathrm{T}}\mu + \frac{1}{2}t^2 u^{\mathrm{T}}\Omega u)$. The MGF of Y is $M_Y(u) = \mathrm{E}(e^{u^{\mathrm{T}}Y}) = M_{u^{\mathrm{T}}Y}(1) = \exp(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u)$, for any $u \in \mathbb{R}^p$, as stated.

(c) Without loss of generality, let $Y_{\mathcal{A}} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$, for $1 \le q < n$, and partition $t^{\mathrm{T}} = (t_{\mathcal{A}}^{\mathrm{T}}, t_{\mathcal{B}}^{\mathrm{T}})$, $\mu^{\mathrm{T}} = (\mu_{\mathcal{A}}^{\mathrm{T}}, \mu_{\mathcal{B}}^{\mathrm{T}})$, etc. Also without loss of generality suppose that $\mathcal{A} \cup \mathcal{B} = \{1, \ldots, n\}$, since otherwise we can just set $t_j = 0$ for $j \notin \mathcal{A} \cup \mathcal{B}$. Then, using matrix algebra, the joint CGF of Y can be written as

$$K_Y(t) = t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t = t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{A}}t_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}\mathcal{B}}t_{\mathcal{B}} + t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{B}}t_{\mathcal{B}}.$$

This equals the sum of the CGFs of $Y_{\mathcal{A}}$ and $Y_{\mathcal{B}}$, i.e.,

$$K_{Y_{\mathcal{A}}}(t) + K_{Y_{\mathcal{B}}}(t) = t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{A}}t_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + +\tfrac{1}{2}t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}\mathcal{B}}t_{\mathcal{B}}$$

if and only if the final term of $K_Y(t)$ equals zero for all t, which occurs if and only if $\Omega_{\mathcal{A}\mathcal{B}} = 0$. Hence the elements of the variance matrix corresponding to $\mathrm{cov}(Y_r, Y_s)$ must equal zero for any $r \in \mathcal{A}$ and $s \notin \mathcal{A}$, as required. Clearly this also holds if $\mathcal{A} \cup \mathcal{B} \neq \{1, \ldots, p\}$.

(d) Each Y_j has mean μ and variance σ^2, and since they are independent, $\mathrm{cov}(Y_j, Y_k) = 0$ for $j \neq k$. If $u \in \mathbb{R}^n$, then $u^{\mathrm{T}}Y$ is a linear combination of normal variables, with mean and variance

$$\sum_{j=1}^n u_j \mu = u^{\mathrm{T}}\mu 1_n, \quad \sum_{j=1}^n u_j^2 \sigma^2 = u^{\mathrm{T}}\sigma^2 I_n u,$$

so $Y \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, as required.

(e) The MGF of $a + BY$ equals

$$\begin{aligned}
\mathrm{E}\left[\exp\{t^{\mathrm{T}}(a + BY)\}\right] &= \mathrm{E}\left[\exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}Y)\}\right] \\
&= e^{t^{\mathrm{T}}a} M_Y(B^{\mathrm{T}}t) \\
&= \exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}\mu + \tfrac{1}{2}(B^{\mathrm{T}}t)^{\mathrm{T}}\Omega(B^{\mathrm{T}}t)\} \\
&= \exp\left\{t^{\mathrm{T}}(a + B\mu) + \tfrac{1}{2}t^{\mathrm{T}}(B\Omega B^{\mathrm{T}})t\right\},
\end{aligned}$$

which is the MGF of the $\mathcal{N}_m(a + B\mu, B\Omega B^{\mathrm{T}})$ distribution.

χ^2 distribution

Definition 5 *If* $Y_j \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$*, then* $W = Y_1^2 + \cdots + Y_\nu^2$ *has the* **non-central chi-square distribution with ν degrees of freedom (df) and non-centrality parameter** $\delta^2 = (\mu_1^2 + \cdots + \mu_n^2)/\sigma^2$*; we write* $W \sim \sigma^2\chi_\nu^2(\delta^2)$*. Then*

$$M_W(t) = \exp\left(\frac{t\sigma^2\delta^2}{1 - 2t\sigma^2}\right)(1 - 2\sigma^2 t)^{-\nu/2}, \quad t < 1/(2\sigma^2).$$

If $\delta^2 = 0$ *and* $\sigma^2 = 1$ *then* W *has the (central)* **chi-square distribution with ν df**, *we write* $W \sim \chi_\nu^2$*, and its* p*-quantile is written* $c_\nu(p)$*.*

Chi-square variables satisfy

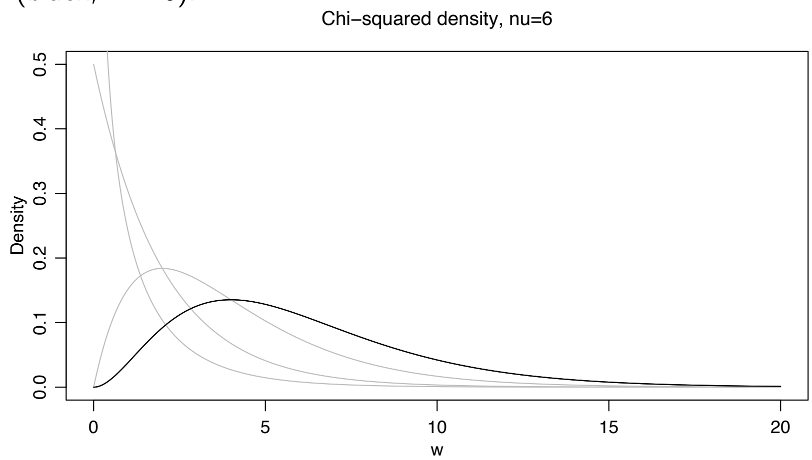
- $\mathrm{E}(W) = \sigma^2(\nu + \delta^2)$, $\mathrm{var}(W) = 2\sigma^4(\nu + 2\delta^2)$;
- if $W_1 \sim \chi_{\nu_1}^2 \perp\!\!\!\perp W_2 \sim \chi_{\nu_2}^2$, then $W_1 + W_2 \sim \chi_{\nu_1+\nu_2}^2$;
- $W \sim \chi_\nu^2$ implies that W has the gamma density

$$f(w) = \frac{\beta^\alpha w^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta w}, \quad w > 0, \quad \alpha, \beta > 0,$$

with $\alpha = \nu/2$ and $\beta = 1/2$.

Central χ_ν^2 densities

Here $\nu = 1, 2, 4, 6$ (black, $\nu = 6$):



Student t distribution

Definition 6 *If $Z \sim \mathcal{N}(0,1) \perp\!\!\!\perp W \sim \chi^2_\nu$, then $T = Z/(W/\nu)^{1/2}$ has the* **Student t distribution with ν df**, *$T \sim t_\nu$, and we write $t_\nu(p)$ for the corresponding p-quantile. The density function of T is*

$$f_T(t) = \frac{\Gamma\{(\nu+1)/2\}}{\sqrt{\nu\pi}\Gamma(\nu/2)} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \quad -\infty < t < \infty, \ \nu = 1, 2, \ldots.$$

Properties:

- the mean and variance exist only for $\nu \geq 2$ et $\nu \geq 3$ respectively, and then

$$\mathrm{E}(T) = 0, \quad \mathrm{var}(T) = \frac{\nu}{\nu - 2};$$

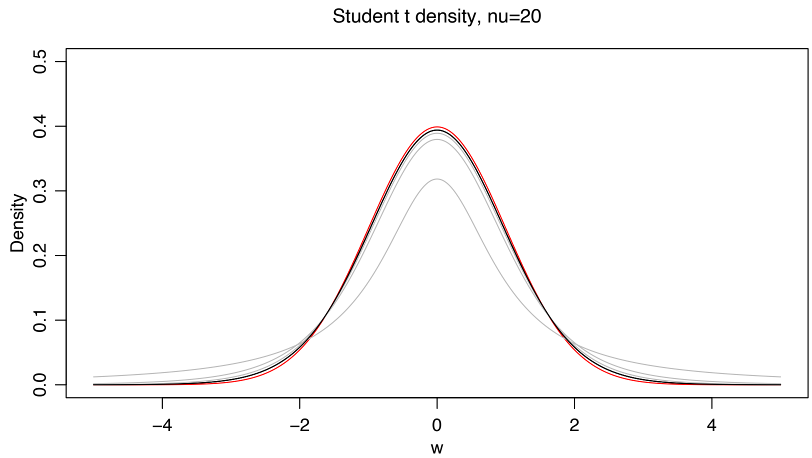
- with $\nu = 1$ we have the **Cauchy density**,

$$\frac{1}{\pi(1+t^2)}, \quad -\infty < t < \infty,$$

  and then T has no moments;
- as $\nu \to \infty$, the limiting distribution of T is $\mathcal{N}(0,1)$; usually the approximation is 'good enough' for $\nu > 25$ (say).

Student t densities

Student t density functions with $\nu = 1, 5, 10, 20$ (black, $\nu = 20$), and the standard normal density (red):



F distribution

Definition 7 *If* $W_1, W_2 \overset{\text{ind}}{\sim} \chi^2_{\nu_1}, \chi^2_{\nu_2}$, *then*

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has the **F distribution with ν_1 and ν_2 df***: we write* $F \sim F_{\nu_1,\nu_2}$.
The density function is

$$f_F(u) = \frac{\Gamma\left(\frac{1}{2}\nu_1 + \frac{1}{2}\nu_2\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} \frac{u^{\frac{1}{2}\nu_1 - 1}}{(\nu_2 + \nu_1 u)^{(\nu_1+\nu_2)/2}}, \quad u > 0,\ \nu_1, \nu_2 = 1, 2, \ldots,$$

and the p-quantile is written $F_{\nu_1,\nu_2}(p)$.

Computation

- ☐ Quantiles of the $\mathcal{N}(\mu, \sigma^2)$, χ^2_ν, t_ν, F_{ν_1,ν_2} distributions can be found in tables, or in environments such as R (see `http://www.r-project.org/`), where they can also be simulated.
- ☐ Examples:

```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.2.1  (2005-12-20 r36812)
...
> qnorm(0.025)     # this is a comment; access normal quantiles
[1] -1.959964      # the [1] means this is the first  element of a vector
> ?qnorm           # help on use of function qnorm()
> qchisq(0.025,df=3)  # chi-squared quantiles, nu=3
[1] 0.2157953
> qt(0.025,df=3)      # t quantiles, nu=3
[1] -3.182446
> qf(0.025,df1=3,df2=4)  # F quantiles, nu1=3, nu2=4
[1] 0.06622087
```

Statistical models

- ☐ Everything above gives a deterministic description of the variation in some numbers y as a linear function of some other numbers X.
- ☐ A **statistical model** is a description of data via a probability distribution.
- ☐ We distinguish
  - **primary** aspects of a model, which specify what questions we aim to answer, from
  - **secondary** aspects, which complete the model, indicate what analysis might be suitable, and determine the precision of conclusions.
- ☐ Often the primary aspects are embodied in one or more **parameters** of the model.
- ☐ (Almost) all models are **tentative**, and we must check that they are reasonable.

Second-order and normal assumptions

- ☐ Two distributional assumptions in general use for the linear model:
  - **second-order assumptions**,

$$\mathrm{E}(y) = X\beta, \quad \mathrm{var}(y) = \sigma^2 V_{n\times n};$$

  - **normal assumptions**,

$$y \sim \mathcal{N}_n(X\beta, \sigma^2 V),$$

    i.e., y has a multivariate normal distribution with mean vector $X\beta$ and (co)variance matrix $\sigma^2 V$, assumed to be positive definite.
- ☐ X is called the **design matrix**: more later.
- ☐ V is assumed known. Unless stated otherwise we set $V = I_n$, so the y_j are uncorrelated; if normal they are therefore independent.
- ☐ If $V \neq I_n$, then we can perform **weighted least squares (WLS)** estimation, minimising

$$\|y - X\beta\|_V^2 = (y - X\beta)^T W (y - X\beta),$$

  where $W = V^{-1}$ is the **weight matrix**.
- ☐ Above the **linearity** is (usually) primary, whereas the **distributional assumption**, use of weights, ..., are (usually) secondary.

Consequences of second-order assumptions

Lemma 8 *Under the second-order assumptions, $\widehat{\beta}$ is an unbiased estimator of β,*

$$\mathrm{E}(\widehat{\beta}) = \beta, \quad \mathrm{var}(\widehat{\beta}) = \sigma^2(X^{\mathrm{T}}X)^{-1}.$$

and $S^2 = (n-p)^{-1}\|y - \widehat{y}\|^2$ is an unbiased estimator of σ^2.

Theorem 9 (Gauss–Markov) *The least squares estimator $\widehat{\beta}$ has the smallest variance among all estimators $\tilde{\beta} = A_{p\times n}y$; it is the* **best linear unbiased estimator (BLUE)** *of β.*

- The results above also hold under the (stronger) normal assumptions.
- We can write

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \sum_{j=1}^{n}(X^{\mathrm{T}}X)^{-1}x_j y_j = n^{-1}\sum_{j=1}^{n} a_j y_j,$$

say, where $a_1, \ldots, a_n$ are $p \times 1$ vectors. Hence as $n \to \infty$ a central limit theorem will apply under mild conditions on the (a_j, y_j), and thus $\widehat{\beta} \dot{\sim} \mathcal{N}_p\{\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}\}$.

Note to Lemma 8

- Recall that expectation is linear, and that $\mathrm{var}(A_{q\times n}y) = A\mathrm{var}(y)A^{\mathrm{T}}$.
- Set $A_{p\times n} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and note that

$$\begin{aligned}\mathrm{E}(\widehat{\beta}) &= \mathrm{E}(Ay) = A\mathrm{E}(y) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta = \beta,\\ \mathrm{var}(\widehat{\beta}) &= A\mathrm{var}(y)A^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}I_n\sigma^2\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}^{\mathrm{T}} = \sigma^2(X^{\mathrm{T}}X)^{-1}.\end{aligned}$$

- Recall that $\mathrm{var}(y) = \mathrm{E}(yy^{\mathrm{T}}) - \mathrm{E}(y)\mathrm{E}(y)^{\mathrm{T}} = \sigma^2 I_n - X\beta\beta^{\mathrm{T}}X^{\mathrm{T}}$, and note that

$$\|y - \widehat{y}\|^2 = (y-\widehat{y})^{\mathrm{T}}(y-\widehat{y}) = y^{\mathrm{T}}(I_n - H)^{\mathrm{T}}(I_n - H)y = y^{\mathrm{T}}(I_n - H)y = \mathrm{tr}\{(I_n - H)yy^{\mathrm{T}}\}.$$

Hence $\mathrm{E}(\|y - \widehat{y}\|^2)$ equals

$$\mathrm{E}[\mathrm{tr}\{(I_n - H)yy^{\mathrm{T}}\}] = \mathrm{tr}\{(I_n - H)\mathrm{E}(yy^{\mathrm{T}})\} = \mathrm{tr}\{(I_n - H)(\sigma^2 I_n + X\beta\beta^{\mathrm{T}}X^{\mathrm{T}})\} = \sigma^2\mathrm{tr}(I_n - H),$$

because $(I_n - H)X = 0$. Moreover $\mathrm{tr}(I_n) = n$ and

$$\mathrm{tr}(H) = \mathrm{tr}\{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\} = \mathrm{tr}\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\} = \mathrm{tr}(I_p) = p,$$

so

$$\mathrm{E}(\|y - \widehat{y}\|^2) = \sigma^2\mathrm{tr}(I_n - H) = \sigma^2(n-p),$$

as required to show that $\mathrm{E}(S^2) = \sigma^2$.

Note to Theorem 9

- ☐ Let $\tilde\beta$ denote any unbiased estimator of β that is linear in y. Then a $p \times n$ matrix A exists such that $\tilde\beta = Ay$, and unbiasedness implies that $\mathrm{E}(\tilde\beta) = AX\beta = \beta$ for any parameter vector β; this entails $AX = I_p$. Now

$$\begin{aligned}
\mathrm{var}(\tilde\beta) - \mathrm{var}(\widehat\beta) &= A\sigma^2 I_n A^{\mathrm{T}} - \sigma^2 (X^{\mathrm{T}}X)^{-1} \\
&= \sigma^2 \left\{AA^{\mathrm{T}} - AX(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}A^{\mathrm{T}}\right\} \\
&= \sigma^2 A(I_n - H)A^{\mathrm{T}} \\
&= \sigma^2 A(I_n - H)(I_n - H)^{\mathrm{T}}A^{\mathrm{T}}
\end{aligned}$$

  and this $p \times p$ matrix is positive semidefinite. Thus $\widehat\beta$ has smallest variance in finite samples among all linear unbiased estimators of β.
- ☐ Note that this is a finite-sample result that holds for all n and X, not an asymptotic result.

Normal-theory linear model

The following results allow us to perform exact inference for the parameters β and σ^2, and in analysis of variance.

Theorem 10 *Under the normal-theory linear model,*

$$\widehat\beta \sim \mathcal{N}\{\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}\} \quad \perp\!\!\!\perp \quad \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Theorem 11 (Cochran) *Let $Y \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ and suppose that*

$$Y^{\mathrm{T}}Y = \sum_{k=1}^{K} Q_k \quad \textit{with} \quad Q_k = Y^{\mathrm{T}}A_k Y,$$

where the matrices A_k are positive semi-definite with ranks r_k. If $r_1 + \cdots + r_K = n$, then the Q_k are mutually independent, and $Q_k/\sigma^2 \sim \chi^2_{r_k}$.

Theorem 12 *If $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$ and H is symmetric and idempotent, then $y^{\mathrm{T}}Hy \sim \sigma^2\chi^2_p(\delta^2)$, where $p = \mathrm{tr}(H)$ and $\sigma^2\delta^2 = \mu^{\mathrm{T}}H\mu$.*

Note to Theorem 10

- ☐ The first part is easy, because $\widehat{\beta}$ is a linear combination of normal variables so it is normal, and its mean and variance matrix were given by Lemma 8.
- ☐ Likewise the residual $e = y - \widehat{y} = (I - H)y$ is a linear combination of y with mean 0_n and variance $(I - H)\sigma^2$, so $e \sim \mathcal{N}_n\{0_p, (I - H)\sigma^2\}$.
- ☐ As $\mathrm{cov}(\widehat{\beta}, e)$ equals

$$\mathrm{cov}\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, (I - H)y\} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{cov}(y)(I - H)^{\mathrm{T}} = \sigma^2(X^{\mathrm{T}}X)^{-1}\{(I - H)X\}^{\mathrm{T}} = 0,$$

we see that $\widehat{\beta}$ is independent of (any function of) e, and therefore in particular of

$$(n - p)S^2/\sigma^2 = \|y - \widehat{y}\|^2/\sigma^2 = e^{\mathrm{T}}e/\sigma^2.$$

- ☐ The eigenvalues of H are p 1's and $n - p$ 0's, so those of $I - H$ are $n - p$ 1's and p 0's. The spectral decomposition implies that there exists an $n \times n$ orthogonal matrix U such that $I - H = UDU^{\mathrm{T}}$, where $D = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$. Note that $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_n$. Thus $Z = U^{\mathrm{T}}e/\sigma$ has mean vector 0_n and variance matrix

$$\mathrm{var}(Z) = U^{\mathrm{T}}\mathrm{var}(e)U/\sigma^2 = U^{\mathrm{T}}(I - H)\sigma^2U/\sigma^2 = U^{\mathrm{T}}UDU^{\mathrm{T}}U = D,$$

i.e. the $Z_1, \ldots, Z_n$ are independent normal variables, $n - p$ of them have variance 1 and p of them have variance 0 and therefore equal 0 with probability one. Hence, as required,

$$(n - p)S^2/\sigma^2 = e^{\mathrm{T}}e/\sigma^2 = (UZ)^{\mathrm{T}}(UZ) = Z^{\mathrm{T}}U^{\mathrm{T}}UZ = \sum_{j=1}^{n-p} Z_j^2 \sim \chi^2_{n-p}.$$

Note to Theorem 11

- First we prove that for any vector of real numbers $y = (y_1, \dots, y_n)^{\mathrm{T}}$, if
$$Q = y^{\mathrm{T}} y = Q_1 + \cdots + Q_K,$$
where $Q_k = y^{\mathrm{T}} A_k y$ and A_k has rank r_k, then if $r_1 + \cdots + r_K = n$, then there exists an orthogonal matrix U such that, with $z = Uy$, we can write
$$Q_1 = z_1^2 + \cdots + z_{r_1}^2, \quad Q_2 = z_{r_1+1}^2 + \cdots + z_{r_1+r_2}^2, \quad \dots, \quad Q_K = z_{r_1+\cdots+r_{K-1}+1}^2 + \cdots + z_n^2. \quad (1)$$
- First let $K = 2$. If so, then $Q = y^{\mathrm{T}} A_1 y + y^{\mathrm{T}} A_2 y$, and there exists an orthonormal matrix U such that $U^{\mathrm{T}} A_1 U$ is diagonal, with r_1 positive eigenvalues $d_1, \dots, d_{r_1}$, say, and $n - r_1 = r_2$ zero eigenvalues. Without loss of generality we can put the r_1 positive eigenvalues first, and set $z = U^{\mathrm{T}} y$, so $y = Uz$.
- Hence the equation $Q = y^{\mathrm{T}} A_1 y + y^{\mathrm{T}} A_2 y$ becomes
$$y^{\mathrm{T}} y = (Uz)^{\mathrm{T}}(Uz) = z^{\mathrm{T}} z = \sum_{j=1}^{n} z_j^2 = (Uz)^{\mathrm{T}} A_1 U z + (Uz)^{\mathrm{T}} A_2 U z = \sum_{j=1}^{r_1} d_j z_j^2 + z^{\mathrm{T}}(U^{\mathrm{T}} A_2 U) z,$$
which yields
$$\sum_{j=1}^{r_1} (1 - d_j) z_j^2 + \sum_{j=r_1+1}^{n} z_j^2 = z^{\mathrm{T}}(U^{\mathrm{T}} A_2 U) z,$$
and the fact that A_2 has rank $r_2 = n - r_1$ implies that $d_1 = \cdots = d_{r_1} = 1$, $Q_1 = \sum_{j=1}^{r_1} z_j^2$ and $Q_2 = \sum_{j=r_1+1}^{n} z_j^2$.
- This is not a result about random variables, but about quadratic forms of real numbers. Clearly it can be iterated to K quadratic forms, giving (1).
- If $Y_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, then $Z = U^{\mathrm{T}} Y / \sigma \sim \mathcal{N}(0, I_n)$, and as $Q_1, \dots, Q_K$ are sums of independent standard normal variables, they are independent, with $Q_k \sim \chi_{r_k}^2$.

Note to Theorem 12

The spectral decomposition of H is VDV^{T}, where D is diagonal with p 1's and $n - p$ 0's, and $z = V^{\mathrm{T}} y \sim \mathcal{N}_n(V^{\mathrm{T}} \mu, \sigma^2 I_n)$; note that the z_j are independent. Now
$$y^{\mathrm{T}} H y = (V^{\mathrm{T}} y)^{\mathrm{T}} D (V^{\mathrm{T}} y) = \sum_{j=1}^{n} d_j z_j^2 = \sum_{j: d_j = 1} z_j^2,$$
which has a (possibly non-central) χ^2 distribution with $p = \mathrm{tr}(H)$ degrees of freedom, scale parameter σ^2 and
$$\sigma^2 \delta^2 = \sum_{j: d_j = 1} \mathrm{E}(z_j)^2 = \sum_{j=1}^{n} d_j \mathrm{E}(z_j)^2 = (V^{\mathrm{T}} \mu)^{\mathrm{T}} D (V^{\mathrm{T}} \mu) = \mu^{\mathrm{T}} H \mu,$$
as announced.

Inference on β

- Theorem 10 implies that for each $r \in \{1, \dots, p\}$,

$$\frac{\widehat{\beta}_r - \beta_r}{\sigma v_{rr}^{1/2}} = Z_r \sim \mathcal{N}(0,1) \quad \perp\!\!\!\perp \quad (n-p)S^2/\sigma^2 = W \sim \chi^2_{n-p},$$

  where v_{rr} the (r,r) element of $(X^{\mathrm{T}}X)^{-1}$, so

$$\frac{\widehat{\beta}_r - \beta_r}{S v_{rr}^{1/2}} = \frac{Z_r}{\sqrt{W/(n-p)}} \sim t_{n-p}.$$

- Hence we can test the hypothesis that $\beta_r = \beta_r^0$ by comparing $(\widehat{\beta}_r - \beta_r^0)/(S v_{rr}^{1/2})$ to the t_{n-p} distribution, and a $(1-\alpha)$ confidence interval for β_r has limits

$$\widehat{\beta}_r \pm S v_{rr}^{1/2} t_{n-p}(\alpha/2), \quad 0 < \alpha < 1.$$

- Likewise $c^{\mathrm{T}}\widehat{\beta} \sim \mathcal{N}\{c^{\mathrm{T}}\beta, \sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c\}$ for any constant vector $c_{p\times 1}$, so we base inference for the scalar $\gamma = c^{\mathrm{T}}\beta$ on

$$\frac{c^{\mathrm{T}}\widehat{\beta} - \gamma}{S\{c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c\}^{1/2}} \sim t_{n-p}.$$

Prediction

- Inference for the value of a further random variable Y_+ with known $p \times 1$ covariate vector x_+ and satisfying the linear model, so $Y_+ \sim \mathcal{N}(x_+^{\mathrm{T}}\beta, \sigma^2)$ independent of the other variables, is performed by noting that $Y_+ \perp\!\!\!\perp \widehat{\beta}, S^2$ and

$$Y_+ - x_+^{\mathrm{T}}\widehat{\beta} \sim \mathcal{N}\left[0, \sigma^2\{1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+\}\right],$$

  so

$$\frac{Y_+ - x_+^{\mathrm{T}}\widehat{\beta}}{S\{1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+\}^{1/2}} \sim t_{n-p},$$

  which leads to prediction intervals for Y_+ once $\widehat{\beta}$ and S have been observed.
- Although we expect inferences for β and σ^2 to hold as approximations under second-order assumptions, this is not the case for inference on Y_+. (Why not?)

Analysis of variance

- The models with mean vectors $X'\beta'$ and $X\beta$, where the columns of X include those of X', give rise to hat matrices H' and H with respective ranks q and p (and $p > q$).
- We can write the response vector and the corresponding **sum of squares** as

$$\begin{aligned} y &= \widehat{y}' + (\widehat{y} - \widehat{y}') + (y - \widehat{y}) = H'y + (H - H')y + (I - H)y, \\ \|y\|^2 &= \|\widehat{y}'\|^2 + \|\widehat{y} - \widehat{y}'\|^2 + \|y - \widehat{y}\|^2. \end{aligned}$$

- Clearly the ranks of H', $H - H'$ and $I - H$ are respectively q, $p - q$ and $n - p$, so Cochran's theorem applies, and if $\beta = 0$, then

$$\|\widehat{y}'\|^2/\sigma^2 \sim \chi^2_q \quad \perp\!\!\!\perp \quad \|\widehat{y} - \widehat{y}'\|^2/\sigma^2 \sim \chi^2_{p-q} \quad \perp\!\!\!\perp \quad \|y - \widehat{y}\|^2/\sigma^2 \sim \chi^2_{n-p},$$

  and therefore (for example) the ratio of sums of squares satisifies

$$\frac{\|\widehat{y} - \widehat{y}'\|^2/(p-q)}{\|y - \widehat{y}\|^2/(n-p)} \sim F_{p-q,n-p},$$

  with the numerator having a non-central χ^2 distribution if $X'\beta' \neq X\beta$.
- This gives a basis for saying whether the reduction in sum of squares $\|\widehat{y} - \widehat{y}'\|^2$ due to augmenting X' to X can be attributable to chance, or if it is 'significant'.

Example 13 *Work out the sums of squares in detail when* $Y_j \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Note to Example 13

- We have $y \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, so $X = 1_n$ and $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = n^{-1}1_n 1_n^{\mathrm{T}}$ clearly has rank $p = 1$. Also $\widehat{y} = Hy = \overline{y}1_n$ and $(I_n - H)1_n = 0$.
- The variable $y' = y - \mu 1_n \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$, so Cochran's theorem applied to y' gives

$$y'^{\mathrm{T}}Hy'/\sigma^2 \sim \chi^2_1 \quad \perp\!\!\!\perp \quad y'^{\mathrm{T}}(I_n - H)y'/\sigma^2 \sim \chi^2_{n-1}.$$

- Now $(I_n - H)y' = (I_n - H)(y - \mu 1_n) = (I_n - H)y = y - \overline{y}1_n$, so

$$y'^{\mathrm{T}}(I_n - H)y' = y^{\mathrm{T}}(I_n - H)y = \sigma^2 V_2 \sim \sigma^2\chi^2_{n-1},$$

  whereas $Hy \sim \mathcal{N}(H\mu 1_n, \sigma^2 HH^{\mathrm{T}}) \sim \mathcal{N}(\mu 1_n, \sigma^2 H)$, and Theorem 12 implies that

$$y^{\mathrm{T}}Hy = \sigma^2 V_1 \sim \sigma^2\chi^2_1(\delta^2),$$

  with $\sigma^2\delta^2 = \mu 1_n^{\mathrm{T}} H \mu 1_n = n\mu^2$. Hence the F statistic we compute is

$$F = \frac{y^{\mathrm{T}}Hy/1}{y^{\mathrm{T}}(I_n - H)y/(n-1)} \overset{\mathrm{D}}{=} \frac{\sigma^2 V_1/1}{\sigma^2 V_2/(n-1)} = \frac{V_1}{V_2/(n-1)},$$

  where V_1 is non-central χ^2 with 1 df and $\delta^2 = n\mu^2$ independent of V_2, which is central chi-squared with $n - 1$ df.
- In fact F is the square of the t statistic for testing $\mu = 0$.

Terms

- In practice we (almost) always include a constant column 1_n in the design matrix and write

$$X\beta = \begin{pmatrix} 1_n & X_1 & \cdots & X_m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} = 1_n\beta_0 + X_1\beta_1 + \cdots + X_m\beta_m,$$

  where the matrices $X_1, \ldots, X_m$, the **terms**, are successively included.
- The model with only 1_n is treated as the baseline model, and has fitted value and residual vector

$$\widehat{y}_0 = \overline{y}1_n, \quad y - \widehat{y}_0 = y - \overline{y}1_n.$$

- Starting from here we ask which terms lead to significant improvements in the fit, as measured by a reduction in the sum of squares.
- The successive degrees of freedom, i.e., the ranks of the corresponding matrices $I - H$, are

$$n - 1 = \nu_0 \geq \nu_1 \geq \cdots \geq \nu_m,$$

  and $\nu_{r+1} = \nu_r$ when the columns of X_{r+1} depend linearly on those of $1_n, X_1, \ldots, X_r$, so inclusion of X_{r+1} does not change the fitted value or improve the fit.

Sums of squares

- Decomposition of baseline residual into orthogonal vectors

$$y - \widehat{y}_0 = (y - \widehat{y}_m) + (\widehat{y}_m - \widehat{y}_{m-1}) + \cdots + (\widehat{y}_1 - \widehat{y}_0)$$

  with $(y - \widehat{y}_0)^{\mathrm{T}}(y - \widehat{y}_0)$ equal to (Pythagoras)

$$(y - \widehat{y}_m)^{\mathrm{T}}(y - \widehat{y}_m) + (\widehat{y}_m - \widehat{y}_{m-1})^{\mathrm{T}}(\widehat{y}_m - \widehat{y}_{m-1}) + \cdots + (\widehat{y}_1 - \widehat{y}_0)^{\mathrm{T}}(\widehat{y}_1 - \widehat{y}_0),$$

  giving **sums of squares decomposition**

$$SS_0 = SS_m + (SS_{m-1} - SS_m) + \cdots + (SS_0 - SS_1),$$

  where
  - SS_r is residual sum of squares for the model with $1_n, X_1, \ldots, X_r$, on ν_r df,
  - $SS_r - SS_{r+1}$ is reduction in model sum of squares due to adding X_{r+1}.
- Normality/geometry: $\widehat{y}_r - \widehat{y}_{r-1}, y - \widehat{y}_m$ are orthogonal linear functions of the data, so SS_m and all the $SS_{r-1} - SS_r$ are mutually independent and Cochran's theorem can be applied.

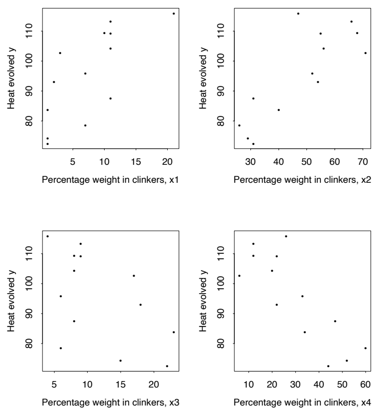
ANOVA table

Terms	df	Residual SS	Terms added	df	Reduction in SS	Mean square
1_n	$n-1$	SS_0				
$1_n, X_1$	ν_1	SS_1	X_1	$n-1-\nu_1$	$SS_0 - SS_1$	$\frac{SS_0 - SS_1}{n-1-\nu_1}$
$1_n, X_1, X_2$	ν_2	SS_2	X_2	$\nu_1 - \nu_2$	$SS_1 - SS_2$	$\frac{SS_1 - SS_2}{\nu_1 - \nu_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$1_n, X_1, \ldots, X_m$	ν_m	SS_m	X_m	$\nu_{m-1} - \nu_m$	$SS_{m-1} - SS_m$	$\frac{SS_{m-1} - SS_m}{\nu_{m-1} - \nu_m}$

- ☐ Usually show only right-hand side and the bottom line of the left-hand side, as 'Residual'.
- ☐ F-tests used to assess need for terms, using reduction in sums of squares relative to the residual estimate of error, SS_m/ν_m.
- ☐ Used to screen which terms give the largest reductions, don't necessarily remove terms that are 'not significant'.

Example: Cement data

Percentage weights in clinkers of 4 four constitutents of cement $(x_1, \ldots, x_4)$ and heat evolved y in calories, in $n = 13$ samples.



Example: Cement data

```
> cement
   x1 x2 x3 x4     y
1   7 26  6 60  78.5
2   1 29 15 52  74.3
3  11 56  8 20 104.3
4  11 31  8 47  87.6
5   7 52  6 33  95.9
6  11 55  9 22 109.2
7   3 71 17  6 102.7
8   1 31 22 44  72.5
9   2 54 18 22  93.1
10 21 47  4 26 115.9
11  1 40 23 34  83.8
12 11 66  9 12 113.3
13 10 68  8 12 109.4
```


Example: Cement data

- ☐ Reductions in overall sum of squares when terms entered in the order given.
- ☐ Clearly x_1 and x_2 should be included, maybe not the others.

Term	df	Reduction in sum of squares	Mean square	F
x_1	1	1450.1	1450.1	242.5
x_2	1	1207.8	1207.8	202.0
x_3	1	9.79	9.79	1.64
x_4	1	0.25	0.25	0.04
Residual	8	47.86	5.98	

Example: Cement data

- ☐ What if we change the order of the terms?

Term	df	Reduction in sum of squares	Mean square	F
x_4	1	1831.9	1831.9	306.2
x_3	1	708.1	708.1	118.4
x_2	1	101.9	101.9	17.04
x_1	1	26.0	26.0	4.34
Residual	8	47.86	5.98	

- ☐ Should x_1 and x_2 be included or not?

Orthogonality

- ☐ In general, ANOVA table depends on order of inclusion of terms.
- ☐ Interpretation unclear if term X_r significant when included early, but not when included late. Is X_r important or not?
- ☐ In a model with orthogonal terms,

$$X\beta = 1_n\beta_0 + X_1\beta_1 + X_2\beta_2, \quad X_r^{\mathrm{T}}X_s = X_r^{\mathrm{T}}1_n = 0, \quad r \neq s.$$

we obtain

$$\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \widehat{\beta}_2\end{pmatrix} = \begin{pmatrix}1^{\mathrm{T}}1 & 0 & 0\\ 0 & X_1^{\mathrm{T}}X_1 & 0\\ 0 & 0 & X_2^{\mathrm{T}}X_2\end{pmatrix}^{-1} \begin{pmatrix}1 & X_1 & X_2\end{pmatrix}^{\mathrm{T}} y$$

so since $\widehat{y} = X\widehat{\beta}$, we have

$$y^{\mathrm{T}}y - \widehat{y}^{\mathrm{T}}\widehat{y} = y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1\widehat{\beta}_1 - \widehat{\beta}_2^{\mathrm{T}}X_2^{\mathrm{T}}X_2\widehat{\beta}_2,$$

and the residual sums of squares for the sub-models $1_n\beta_0$, $1_n\beta_0 + X_1\beta_1$, $1_n\beta_0 + X_2\beta_2$ are

$$y^{\mathrm{T}}y - n\overline{y}^2, \quad y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1\widehat{\beta}_1, \quad y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_2^{\mathrm{T}}X_2^{\mathrm{T}}X_2\widehat{\beta}_2,$$

so clearly the reductions do not depend on the order of inclusion.

Balance

- ☐ Balanced design matrices induce orthogonality after fitting 1_n (or maybe a more complex design).
- ☐ Gram–Schmidt orthogonalisation with respect to early terms makes later terms mutually orthogonal, leading to a clear interpretation of the ANOVA.
- ☐ If we denote the centered versions of X_1 and X_2 as

$$Z_r = (I_n - n^{-1}1_n1_n^{\mathrm{T}})X_r = X_r - 1_n\overline{x}_r^{\mathrm{T}} \quad r = 1, 2,$$

where $\overline{x}_r$ is the row average for X_r, then we can write

$$1_n\beta_0 + X_1\beta_1 + X_2\beta_2 = 1_n(\beta_0 + \overline{x}_1^{\mathrm{T}}\beta_1 + \overline{x}_2^{\mathrm{T}}\beta_2) + Z_1\beta_1 + Z_2\beta_2 = 1_n\gamma_0 + Z_1\beta_1 + Z_2\beta_2,$$

so $Z_1^{\mathrm{T}}1_n = Z_2^{\mathrm{T}}1_n = 0$.
- ☐ If the design is such that $Z_1^{\mathrm{T}}Z_2 = 0$, then the order of inclusion of X_1, X_2 is irrelevant, provided 1_n is included first.

Example 14 (3×2 layout) *Observations and their means written as*

$$\begin{array}{cc} y_{11} & y_{12} \\ y_{21} & y_{22}, \\ y_{31} & y_{32} \end{array} \qquad \begin{array}{cc} \mu & \mu + \alpha \\ \mu + \delta_1 & \mu + \delta_1 + \alpha. \\ \mu + \delta_2 & \mu + \delta_2 + \alpha \end{array}$$

Note to Example 14

In terms of the parameter vector $(\mu, \alpha, \delta_2, \delta_3)^{\mathrm{T}}$, the design matrix is

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix},$$

with X_1 the second column of X, and X_2 the third and fourth columns of X. Evidently X_1 and X_2 are not orthogonal and they are not orthogonal to 1_n. On the other hand Z_1 and Z_2 in the corresponding centred matrix,

$$\begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{3} & -\frac{1}{3} \\ 1 & \frac{1}{2} & -\frac{1}{3} & -\frac{1}{3} \\ 1 & -\frac{1}{2} & \frac{2}{3} & -\frac{1}{3} \\ 1 & \frac{1}{2} & \frac{2}{3} & -\frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{3} & \frac{2}{3} \\ 1 & \frac{1}{2} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

are orthogonal to the constant by construction and to each other because the design is balanced: δ_2 and δ_3 each occur equally often with α and without α. This balance implies that provided that μ is fitted first, the reductions in sums of squares due to X_1 and X_2, or equivalently Z_1 and Z_2, are unique.

Diagnostics slide 36

Assumptions and model checking

- ☐ How heavily do our conclusions depend on our assumptions?
- ☐ In any given context,
  - **primary** aspects relate to the questions our analysis will address,
  - **secondary** aspects relate to how we go about finding answers to them.
- ☐ Queries about primary aspects suggest that we should start again.
- ☐ Queries about secondary aspects suggest that we modify the analysis.
- ☐ **Regression diagnostics** check that a fitted model is adequate:
  - Does y depend linearly on the columns of X?
  - Does y depend systematically on variables omitted from X?
  - Are the variances constant?
  - Are the responses uncorrelated/independent?
  - Are there outliers or otherwise unusual data?
  - Are the responses normally distributed?
- ☐ Usually these involve plots, sometimes tests — **beware over-interpretation**!
- ☐ Key question: ‘how would the failure I see/suspect change my conclusions?’

Residuals

- **Raw residuals** defined as

$$e = y - \widehat{y} = y - X\widehat{\beta} = (I_n - H)y$$

have $\mathrm{E}(e) = 0$, $\mathrm{var}(e) = \sigma^2(I_n - H)$ if model correct, so

$$\mathrm{var}(e_j) = \sigma^2(1 - h_{jj}) \quad \mathrm{cov}(e_j, e_k) = -\sigma^2 h_{jk},\ j \neq k.$$

In a good model e and $\widehat{y}$ should be unrelated, because $e^{\mathrm{T}}\widehat{y} = 0$.

- Prefer **standardized residuals**

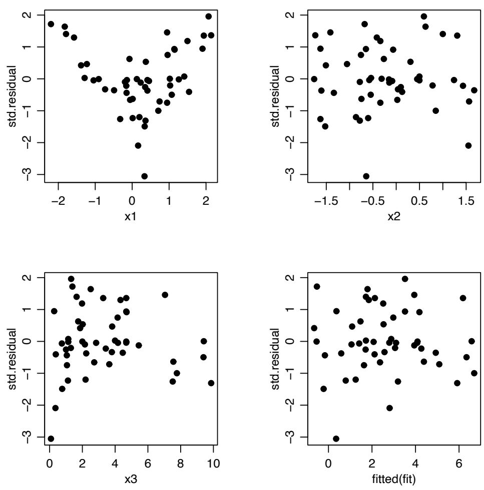
$$r_j = \frac{e_j}{s(1 - h_{jj})^{1/2}} = \frac{y_j - x_j^{\mathrm{T}}\widehat{\beta}}{s(1 - h_{jj})^{1/2}}, \quad j = 1, \ldots, n,$$

which satisfy $\mathrm{E}(r_j) = 0$ and $\mathrm{var}(r_j) \doteq 1$; here s is residual standard deviation.

- Uses:
  - check linearity by plotting r_j against covariates, both those in X and not;
  - check constant variance by plotting r_j against fitted value $\widehat{y}_j$;
  - check independence by ACF of residuals (if data time-ordered);
  - outliers visible as unusual residuals;
  - check normality by normal QQ-plot of r_j.

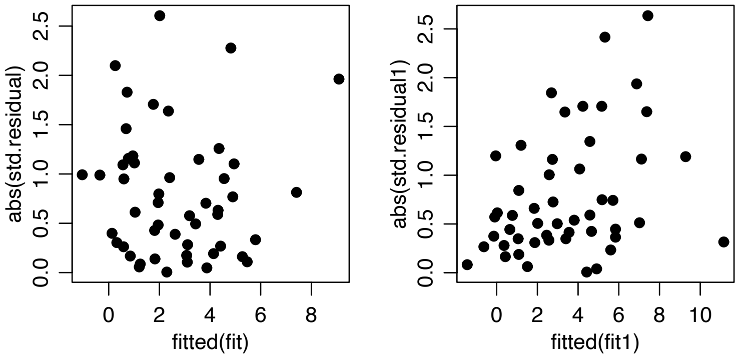
Checking linearity

- Plot r against each covariate, included or not in the model, and against $\widehat{y}$, which is uncorrelated with e (as $\widehat{y}^{\mathrm{T}}e = 0$):



Checking the variance

- ☐ Does $\mathrm{var}(y)$ depend on $\mathrm{E}(y)$?
- ☐ Variance function shows how $\mathrm{var}(y)$ depends on $\mu = \mathrm{E}(y)$. For normal linear model should have $\mathrm{var}(y) = \sigma^2$, so variance is constant function of μ
- ☐ Plot r or $|r|$ against $\widehat{y}$:



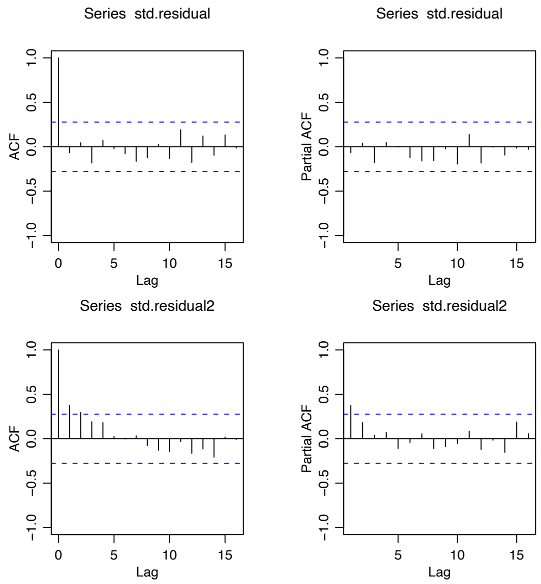
Checking independence

- ☐ Dependence can greatly increase uncertainty of final conclusions.
- ☐ Substantive knowledge is helpful in suggesting whether it might be present:
  - were the data gathered in temporal/spatial/. . . order?
  - were the data sampled/gathered in groups (e.g., spatial, several observations on different individuals, . . .)?
  - was randomisation used? If so, how?
- ☐ If observations are time-ordered, try using correlogram (ACF) and partial correlogram (PACF) to estimate serial correlations and partial correlations

$$\mathrm{corr}(r_j, r_{j+t}), \quad \mathrm{corr}(r_j, r_{j+t} \mid r_{j+1}, \ldots, r_{j+t-1}), \quad t = 1, \ldots$$

- ☐ On next page, top panels show uncorrelated residuals, lower ones show evidence of correlation, suggesting use of a time series model.

Checking independence



Checking normality

- ☐ Normal Q-Q plot for $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ graphs ordered values

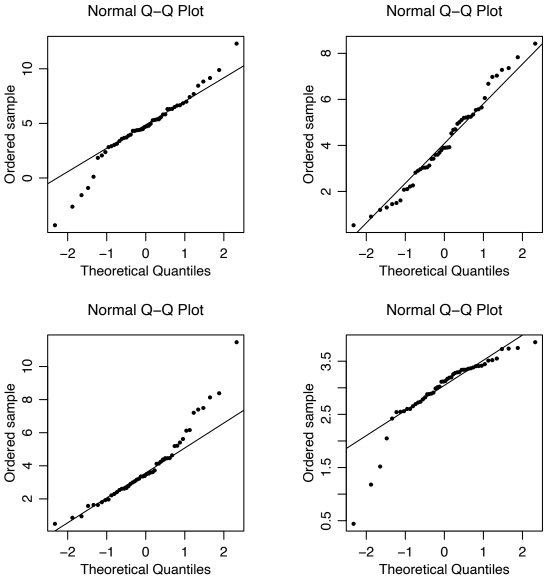
$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

  against (approximate) expected normal order statistics

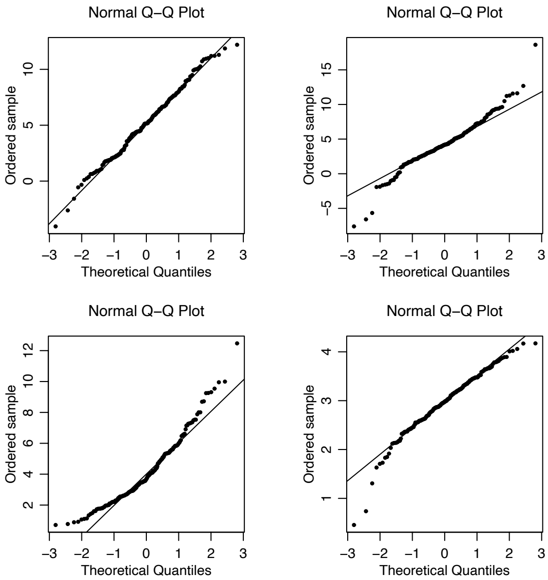
$$\Phi^{-1}\{1/(n+1)\}, \Phi^{-1}\{2/(n+1)\}, \ldots, \Phi^{-1}\{n/(n+1)\}.$$

- ☐ Normality — roughly straight line, slope σ, intercept μ.
- ☐ Outliers, skewness, heavy tails (easily) seen.
- ☐ Beware over-interpretation of such plots when n is small — often useful to add simulation envelope.
- ☐ Apply to standardized residuals r_j from regression model.

Checking normality, $n = 50$



Checking normality, $n = 200$



Leverage and influence

- Does **case** (x_j, y_j) strongly influence the fitted model (picture)?
- As
$$\text{var}(y_j - \widehat{y}_j) = \text{var}(y_j - x_j^{\mathrm{T}}\beta) = s^2(1 - h_{jj}),$$
having **leverage** $h_{jj} \doteq 1$ implies that $\widehat{y}_j \approx y_j$ — need one parameter to fit this case.
- As $\text{tr}(H) = \sum_{j=1}^{n} h_{jj} = p$, the average h_{jj} is p/n. If $h_{jj} > 2p/n$, then jth case should be checked (rule of thumb), e.g. by refitting without (x_j, y_j).
- Let $\widehat{y}_{-j}$ be fitted values for (all) data when (x_j, y_j) is dropped and use **Cook's distance**
$$C_j = \frac{1}{ps^2}(\widehat{y} - \widehat{y}_{-j})^{\mathrm{T}}(\widehat{y} - \widehat{y}_{-j}) = \cdots = \frac{r_j^2 h_{jj}}{p(1 - h_{jj})}$$
to measure the difference between $\widehat{y}$ and $\widehat{y}_{-j}$.
- Large C_j implies large r_j and/or large h_{jj}.
- Cases with $C_j > 8/(n - 2p)$ worth a closer look (rule of thumb).
- High leverage and/or influence need not be bad, just need to be aware of it.
- These ideas are not very useful in large samples, since the plots become uninformative.

Response transformation

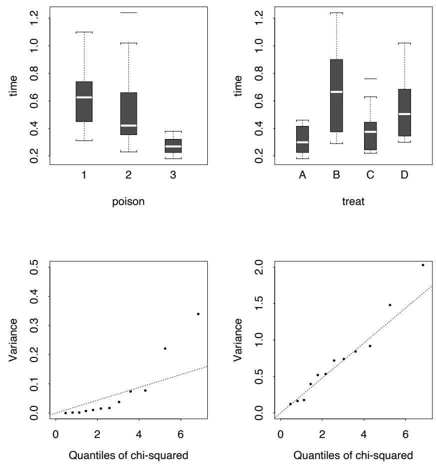
- Linear model for y may be better applied for some transformation $g(y)$, especially if some y are much larger than others, or the variance is non-constant.
- Survival times y_{ptj} in 10-hour units of animals in a 3×4 factorial experiment with four replicates, with (below) average (standard deviation) for the poison $\times$ treatment combinations:
  - generally see higher SD and mean together,
  - times must be positive, so linear model inappropriate?

Treatment	Poison 1	Poison 2	Poison 3
A	0.31, 0.45, 0.46, 0.43	0.36, 0.29, 0.40, 0.23	0.22, 0.21, 0.18, 0.23
B	0.82, 1.10, 0.88, 0.72	0.92, 0.61, 0.49, 1.24	0.30, 0.37, 0.38, 0.29
C	0.43, 0.45, 0.63, 0.76	0.44, 0.35, 0.31, 0.40	0.23, 0.25, 0.24, 0.22
D	0.45, 0.71, 0.66, 0.62	0.56, 1.02, 0.71, 0.38	0.30, 0.36, 0.31, 0.33

Treatment	Poison 1	Poison 2	Poison 3	Average
A	0.41 (0.07)	0.32 (0.08)	0.21 (0.02)	0.31
B	0.88 (0.16)	0.82 (0.34)	0.34 (0.05)	0.68
C	0.57 (0.16)	0.38 (0.06)	0.24 (0.01)	0.39
D	0.61 (0.11)	0.67 (0.27)	0.33 (0.03)	0.53
Average	0.62	0.55	0.28	0.48

Example: Poison data

Upper panels: dependence of responses on the factor levels. Lower left: χ^2_3 probability plots of the $3s^2_{pt}$, where s^2_{pt} is the sample variance of y_{ptj}. Lower right: same for y^{-1}_{ptj}.



Box–Cox transformation

- For $y > 0$, the **Box–Cox transformation**

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0, \end{cases}$$

  includes the inverse ($\lambda = -1$), log ($\lambda = 0$), cube and square roots ($\lambda = \frac{1}{3}, \frac{1}{2}$), original scale ($\lambda = 1$) and square ($\lambda = 2$); sometimes map $y \mapsto y + c > 0$.
- Suppose normal linear model

$$y^{(\lambda)} \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

  applies for some β, σ and λ to be determined. Here X contains 1_n, so use of $y^{(\lambda)}$ just changes intercept and rescales β and σ.
- Use profile log likelihood for λ to choose 'best' transformation (usually from list above to aid interpretation).
- Interpretation of β depends on λ, so usually we ignore the fact that λ was estimated, unless we are not interested in β (e.g., when performing 'automatic' prediction).

Example: Poison data

- Fits of two-way layout model, with interaction:

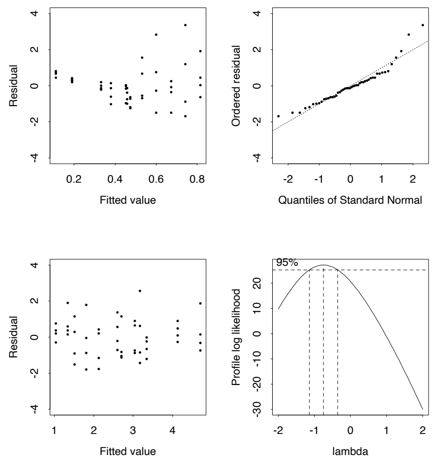
$$y_{tpj} \sim \mathcal{N}(\mu + \alpha_t + \beta_p + \gamma_{tp}, \sigma^2), \quad t = 1,2,3,4,\ p = 1,2,3,\ j = 1,2,3,4.$$

- Analyses of variance with responses y and y^{-1}. For MS and F read 'Mean square' and 'F statistic'.
- The terms explain appreciably more variation for y^{-1}.

Term	df	Response y			Response y^{-1}		
		SS	MS	F	SS	MS	F
Poisons	2	1.033	0.517	23.22	34.88	17.44	72.63
Treatments	3	0.921	0.307	13.81	20.41	6.80	28.34
Treatments × Poisons	6	0.250	0.042	1.87	1.57	0.26	1.09
Residual	36	0.801	0.022		8.64	0.24	

Example: Poison data

Top: residuals for model without interactions γ_{tp}; clearly problematic.
Lower right: profile log likelihood for Box–Cox λ, showing 95% confidence interval.
Lower left: residuals for the two-way layout model (no interactions) for $1/y$.



Goals

- ☐ What to do faced with a set of data?
- ☐ Two main aims:
  - **understand** (science) — maybe have prior idea/hypotheses on how response depends on explanatory variables. Interpretation is key.
  - **predict/control** (technology) — don't really care how y depends on X. Interpretation not critical (though this describes only prediction in the narrowest of senses).
- ☐ There is no reason that a single model will do both, or even that there must be a single 'best' model:
  - maybe two models with different interpretations both fit about equally well, and then future work might aim to choose between them;
  - prediction with a mixture of models might be better than using a single model.

Meta-algorithm

- ☐ **Collect** data intended to answer question of interest;
- ☐ **examine** data (graphs, look for outliers, problems with sampling scheme);
- ☐ **choose/construct** response variable (transformations? independence?);
- ☐ **consider** what models are coherent with context of problem (limiting properties, units, similar problems/datasets, covariates that must be included, ...);
- ☐ **iterate**:
  - fit models, compare quality of fits;
  - check interpretations of $\widehat{\beta}, \widehat{\sigma}^2$ and
  - check fit (diagnostics, outliers, ...)

  until satisfied; finally
- ☐ give **conclusions**—careful interpretation of best model(s) in terms of original problem, consider deficiencies, and explain what extra data might overcome them.

Initial examination of data

- ☐ Plot y against covariates, look for outliers, non-constant variance, nonlinearity, etc.
- ☐ Plot covariates against each other, look for dependence.
- ☐ Try to understand covariates (e.g., dimensions), are transformations needed?
- ☐ May need to reduce dimension of X by **regularisation** — many ways to do this (later).

Albert Einstein (1879–1955)

'Everything should be made as simple as possible, *but no simpler*.'

William of Occam (?1285–1347/9)

Occam's razor: *Pluralitas non est ponenda sine necessitate*: entities should not be multiplied beyond necessity.

Automatic variable selection

- ☐ Assume linear model $\mathrm{E}(y) = X\beta$
- ☐ 2^p possible subsets of columns of X, plus transformations, ...
- ☐ Example: $p = 17$ gives 131072 possible subsets of variables
- ☐ Fast algorithms (e.g., branch and bound, `leaps` in `R`) exist visit them all or just subsets (e.g., stepwise), but we need criteria for comparing models.
- ☐ Many proposals for model comparison
  - cross-validation,
  - information criteria (AIC, AIC$_\mathrm{c}$, BIC, NIC, TIC, ...)
  - Mallow's C_p,
  - ...
- ☐ Most involve minimising estimated prediction error for future data *like those observed*!

Prediction error

- ☐ True model $y \sim (\mu, \sigma^2 I_n)$, we assume (perhaps incorrectly) that $\mu = X\beta$, fit $X_{n\times p}$ and obtain fitted value
$$X\widehat{\beta} = Hy \sim (H\mu, \sigma^2 H).$$
- ☐ Terminology
  - the **true model** has $\mu = X\beta$ and all $\beta_r \neq 0$;
  - a **correct model** has $\mu = X\beta$ but some $\beta_r = 0$;
  - a **wrong** model has $\mu \notin \mathrm{span}(X)$;

  so $(I_n - H)\mu = 0$ if the model is true or correct, and $(I_n - H)\mu \neq 0$ if it is wrong.
- ☐ The **prediction error** for an independent dataset y_+ with mean vector μ is
$$\Delta = n^{-1}\mathrm{E}\left\{(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta})\right\} = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I-H)\mu + (1+p/n)\sigma^2, & \text{wrong},\\ (1+q/n)\sigma^2, & \text{true},\\ (1+p/n)\sigma^2, & \text{correct}, \end{cases}$$
  where $\mathrm{E}(\cdot)$ is over both y_+ and y and $p \geq q = \#\{\beta_r : \beta_r \neq 0\}$ when $\mu \in \mathrm{span}(X)$.
- ☐ In principle we should write $\Delta \equiv \Delta(X)$.

Note: Computation of Δ

Let $y \sim (\mu, \sigma^2 I_n)$ and fit $X\beta$, obtaining fitted value

$$X\widehat{\beta} = Hy \sim (H\mu, \sigma^2 H),$$

where $H\mu = \mu$, i.e., $(I_n - H)\mu = 0$ if $\mu \in \text{span}(X)$, but otherwise $(I_n - H)\mu \neq 0$.
We have a new data set $y_+ \sim (\mu, \sigma^2 I_n)$, and we compute the average error in predicting y_+ using $X\widehat{\beta}$, i.e.,

$$\Delta = n^{-1}\mathrm{E}\left\{(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta})\right\}.$$

Let $A = y_+ - X\widehat{\beta}$ and note that as the trace of a scalar is the scalar and trace is a linear operator,

$$\mathrm{E}(A^{\mathrm{T}}A) = \mathrm{E}\{\mathrm{tr}(A^{\mathrm{T}}A)\} = \mathrm{E}\{\mathrm{tr}(AA^{\mathrm{T}})\} = \mathrm{tr}\{\mathrm{E}(AA^{\mathrm{T}})\} = \mathrm{tr}\{\mathrm{var}(A) + \mathrm{E}(A)\mathrm{E}(A)^{\mathrm{T}}\}.$$

Now as y_+ and y are independent,

$$y_+ - X\widehat{\beta} \sim (\mu - H\mu, \sigma^2 I_n + \sigma^2 H),$$

so the computation above gives

$$\mathrm{E}\left\{(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta})\right\} = \mathrm{tr}\{\sigma^2(I_n + H) + (I_n - H)\mu\mu^{\mathrm{T}}(I_n - H)\} = \sigma^2(n+p) + \mu^{\mathrm{T}}(I_n - H)\mu,$$

because $\mathrm{tr}(I_n + H) = n + p$ and $I - H$ is symmetric and idempotent, giving

$$\Delta = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I - H)\mu + (1 + p/n)\sigma^2, & \text{wrong model}, \\ (1 + q/n)\sigma^2, & \text{true model}, \\ (1 + p/n)\sigma^2, & \text{correct model}. \end{cases}$$

Bias/variance trade-off

- Minimising Δ involves balancing the
  - **bias** $n^{-1}\mu^{\mathrm{T}}(I - H)\mu$, which is reduced by including useful terms in X, and
  - **variance** $(1 + p/n)\sigma^2$, which is increased by including useless terms in X.
- We would like to minimise Δ, but this depends the unknown μ and σ.
- The **cross-validation** estimator of Δ splits the data into X', y' and X^*, y^*, then
  - for each possible subset $\mathcal{S}$ of columns of X^*:
    - compute $\widehat{\beta}^*_{\mathcal{S}}$ by regressing y^* on $X^*_{\mathcal{S}}$;
    - use $\widehat{\beta}^*_{\mathcal{S}}$ to estimate the prediction error for $\mathcal{S}$ by

      $$\widehat{\Delta}_{\mathcal{S}} = (n')^{-1}(y' - X'_{\mathcal{S}}\widehat{\beta}^*_{\mathcal{S}})^{\mathrm{T}}(y' - X'_{\mathcal{S}}\widehat{\beta}^*_{\mathcal{S}});$$

  - finally choose the set of columns $\mathcal{S}$ for which $\widehat{\Delta}_{\mathcal{S}}$ is minimised.
- This estimator depends on the split, and since $X' \neq X^*$ in general, $\widehat{\Delta}_{\mathcal{S}}$ does not estimate $\Delta_{\mathcal{S}}$.

Leave-one-out cross-validation

- Simplest way to use entire dataset is **leave-one-out cross-validation (CV)**, minimising

$$n\widehat{\Delta}_{\mathrm{CV}} = \mathrm{CV} = \sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j})^2,$$

where $\widehat{\beta}_{-j}$ is estimate computed without (x_j, y_j). This seems to require n fits of each model, but (exercise)

$$\mathrm{CV} = \sum_{j=1}^{n}\frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2}{(1-h_{jj})^2},$$

which can be obtained from one fit.

- **Generalised cross-validation (GCV)** replaces h_{jj} by its average $\mathrm{tr}(H)/n = p/n$, giving

$$\mathrm{GCV} = \sum_{j=1}^{n}\frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2}{(1-p/n)^2},$$

and hence

$$\mathrm{E}(\mathrm{GCV}) = \mu^{\mathrm{T}}(I-H)\mu/(1-p/n)^2 + n\sigma^2/(1-p/n) \approx n\Delta.$$

- Often choose the model that minimises GCV or CV.

Note: Properties of GCV

We need the expectation of $\varepsilon^{\mathrm{T}}\varepsilon$, where $\varepsilon = y - X\widehat{\beta} = (I-H)y \sim ((I_n - H)\mu, (I_n - H)\sigma^2)$, and

$$\mathrm{E}\left(\varepsilon^{\mathrm{T}}\varepsilon\right) = \mathrm{E}\left\{\mathrm{tr}\left(\varepsilon\varepsilon^{\mathrm{T}}\right)\right\} = \mathrm{tr}\left\{\mathrm{E}(\varepsilon)\mathrm{E}(\varepsilon)^{\mathrm{T}} + \mathrm{var}(\varepsilon)\right\} = \mu^{\mathrm{T}}(I-H)\mu + \sigma^2\mathrm{tr}(I_n - H).$$

Now note that $\mathrm{tr}(H) = p$ and divide by $(1-p/n)^2$ to give (almost) the required result, for which we need also $(1-p/n)^{-1} \approx 1 + p/n$, for $p \ll n$.

Akaike information criterion

- The above arguments apply only to least squares estimators. More generally, we could aim to minimise the **Kullback–Leibler discrepancy (information distance)**

$$D(f_\theta, g) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)\,\mathrm{d}y \ge 0,$$

  between **candidate model** $f_\theta \equiv f(y;\theta)$ and true model g, based on $Y_1,\ldots,Y_n \stackrel{\text{iid}}{\sim} g$.
- This fails, because many candidates may have the same minimum, so we need to penalise the dimension of θ.
- Suppose that θ_g minimises $D(f_\theta, g)$ within the family of candidate models, and is estimated from $Y_1,\ldots,Y_n$ using the MLE $\widehat\theta$.
- We suppose there is an independent sample $Y_1^+,\ldots,Y_n^+ \stackrel{\text{iid}}{\sim} g$ and aim to estimate

$$\mathrm{E}_g\left(\mathrm{E}_g^+\left[\sum_{j=1}^n \log\left\{\frac{g(Y_j^+)}{f(Y_j^+;\widehat\theta)}\right\}\right]\right) = n\mathrm{E}_g\left\{D(f_{\widehat\theta}, g)\right\}; \qquad (2)$$

  the outer expectation is over the distribution of $\widehat\theta$, which is independent of Y^+. Having $\dim\theta$ too high allows $\widehat\theta$ to miss θ_g by more, which should penalise $D(f_{\widehat\theta}, g)$.
- We need to know the distribution of $\widehat\theta$ in a mis-specified model.

Likelihood asymptotics

- In a regular model where $p = \dim(\theta)$ and $Y_1,\ldots,Y_n \stackrel{\text{iid}}{\sim} g \equiv f_{\theta_g}$, then

$$\widehat\theta \,\dot\sim\, \mathcal{N}_p\left\{\theta_g, n^{-1}I_g(\theta_g)^{-1}\right\}, \quad I_g(\theta) = -\int \frac{\partial^2 \log f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\, g(y)\,\mathrm{d}y,$$

  but in the mis-specified case $f_{\theta_g} \ne g$ and (under regularity conditions) we obtain

$$\widehat\theta \,\dot\sim\, \mathcal{N}_p\left\{\theta_g, n^{-1}I(\theta_g)^{-1}K(\theta_g)I(\theta_g)^{-1}\right\}, \quad K(\theta) = \int \frac{\partial \log f(y;\theta)}{\partial\theta}\frac{\partial \log f(y;\theta)}{\partial\theta^{\mathrm{T}}}\, g(y)\,\mathrm{d}y.$$

- Taylor expansions lead to

$$n\mathrm{E}_g\left\{D(f_{\widehat\theta}, g)\right\} \doteq nD(f_{\theta_g}, g) + \tfrac{1}{2}\mathrm{tr}\left\{I_g(\theta_g)^{-1}K(\theta_g)\right\},$$

  and the second term equals $p/2$ in the regular case, since then $I_g(\theta_g) = K(\theta_g)$.
- To estimate $n\mathrm{E}_g\left\{D(f_{\widehat\theta}, g)\right\}$ we ignore the terms with g only and after more expansions the estimator is the **Akaike information criterion**

$$\mathrm{AIC} = -2(\widehat\ell - p) \qquad (\equiv n\log\mathrm{RSS} + 2p \text{ in linear model})$$

  where $\widehat\ell$ is the maximised log likelihood for the model f_θ.

Note: Derivation of AIC

- ☐ Taylor series expansion shows that $\log f(y;\widehat{\theta})$ approximately equals

$$\log f(y;\theta_g) + (\widehat{\theta}-\theta_g)^{\mathrm{T}}\frac{\partial \log f(y;\theta_g)}{\partial\theta} + \tfrac{1}{2}(\widehat{\theta}-\theta_g)^{\mathrm{T}}\frac{\partial^2 \log f(y;\theta_g)}{\partial\theta\partial\theta^{\mathrm{T}}}(\widehat{\theta}-\theta_g),$$

and as θ_g minimizes $D(f_\theta, g)$,

$$\int \frac{\partial \log f(y;\theta_g)}{\partial\theta} g(y)\,dy = 0.$$

Hence taking expectation over $Y_1^+,\ldots,Y_n^+$, we get

$$nD(f_{\widehat{\theta}}, g) = n\int \log\left\{\frac{g(y)}{f(y;\widehat{\theta})}\right\} g(y)\,dy \doteq nD(f_{\theta_g}, g) + \tfrac{1}{2}\mathrm{tr}\left\{(\widehat{\theta}-\theta_g)(\widehat{\theta}-\theta_g)^{\mathrm{T}} I_g(\theta_g)\right\},$$

where we have used the fact that the trace of a scalar is itself.

- ☐ Expectation over the distribution of $\widehat{\theta}$ gives its variance matrix, $I_g(\theta_g)^{-1}K(\theta_g)I_g(\theta_g)^{-1}$, and hence

$$nE_g\left\{D(f_{\widehat{\theta}}, g)\right\} \doteq nD(f_{\theta_g}, g) + \tfrac{1}{2}\mathrm{tr}\left\{I_g(\theta_g)^{-1}K(\theta_g)\right\}, \quad (3)$$

where the second term penalizes the dimension p of θ. The first term here is $O(n)$ but the second is $O(p)$. When $f_{\theta_g} = g$, $I_g(\theta_g) = K(\theta_g)$ so $\mathrm{tr}\left\{I_g(\theta_g)^{-1}K(\theta_g)\right\} = p$.

- ☐ To build an estimator, note that $\int \log g(y)\, g(y)\, dy$ is constant and can be ignored. Now $\ell(\widehat{\theta}) = \ell(\theta_g) + \{\ell(\widehat{\theta}) - \ell(\theta_g)\}$, so

$$\begin{aligned} E_g\left\{-\ell(\widehat{\theta})\right\} &= -E_g\left\{\ell(\theta_g) + \tfrac{1}{2}W(\theta_g)\right\} \\ &\doteq nD(f_{\theta_g}, g) - \tfrac{1}{2}\mathrm{tr}\left\{I(\theta_g)^{-1}K(\theta_g)\right\} - n\int \log g(y)\, g(y)\, dy, \end{aligned}$$

where we have used the fact that under the wrong model, the likelihood ratio statistic $W(\theta_g)$ has mean approximately $\mathrm{tr}\left\{I(\theta_g)^{-1}K(\theta_g)\right\}$. Hence $-\ell(\widehat{\theta})$ tends to underestimate $nD(f_{\theta_g}, g) - n\int \log g(y)\, g(y)\, dy$. On reflection this is obvious, because $\ell(\widehat{\theta}) \geq \ell(\theta_g)$ by definition of $\widehat{\theta}$. As p increases, so will the extent of overestimation.

- ☐ An estimator is $-\ell(\widehat{\theta}) + c$, where c estimates $\mathrm{tr}\left\{I(\theta_g)^{-1}K(\theta_g)\right\}$. Two possible choices of c are p and $\mathrm{tr}(\widehat{I}^{-1}\widehat{K})$, and these lead to

$$\mathrm{AIC} = 2\{-\ell(\widehat{\theta}) + p\}, \quad \mathrm{NIC} = 2\{-\ell(\widehat{\theta}) + \mathrm{tr}(\widehat{J}^{-1}\widehat{K})\}; \quad (4)$$

another possibility is $\mathrm{BIC} = -2\ell(\widehat{\theta}) + p\log n$.

- ☐ The model is chosen to minimize AIC, say, with the factor 2 putting differences of AIC on the same scale as likelihood ratio statistics. Such criteria are used far beyond random samples, and even in cases where the theory above doesn't work.
- ☐ In particular, the maximised log-likelihood for a normal-theory linear model with residual sum of squares RSS can be shown to be

$$-\frac{n}{2}\log(2\pi\widehat{\sigma}) - \frac{n}{2} \equiv -\frac{n}{2}\log \mathrm{RSS} + \text{constants},$$

which leads to the formula given on the slide.

Other model selection criteria

- ☐ 'Corrected' AIC for (normal-theory) regression problems:

$$\mathrm{AIC_c} \equiv n \log \widehat{\sigma}^2 + n \frac{1 + p/n}{1 - (p+2)/n}.$$

- ☐ Also can use **Bayes' information criterion**

$$\mathrm{BIC} = -2\widehat{\ell} + p \log n.$$

- ☐ Mallows suggested

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

  where SS_p is RSS for fitted model and s^2 estimates σ^2.
- ☐ When the true model is a candidate and $n \to \infty$,
  - AIC is **inconsistent** — it will not choose the true model with probability one, but tends to pick a more complex model;
  - $\mathrm{AIC_c}$ is also inconsistent but gives better results in finite samples;
  - BIC is **consistent** — it chooses the true model with probability $\to 1$.

  These results suppose that the models are fixed, but in practice we also have $p \to \infty$ when $n \to \infty$, because we fit ever more complex models when we have more data.

Simulation experiment

Number of times models were selected using various model selection criteria in 50 repetitions using simulated normal data for each of 20 design matrices. The true model has $p = 3$.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	$\mathrm{AIC_c}$	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	$\mathrm{AIC_c}$		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	$\mathrm{AIC_c}$			786	105	52	41	16

Stepwise methods

- ☐ In principle we might wish to fit all 2^p possible choices of covariates, but in practice this is possible only for 'modest' p, using **leaps** or similar methods (or approximations).
- ☐ When p is too large for exhaustive searches, we instead consider subsets of the models, using the methods below (or variants).
- ☐ **Forward selection**: starting from model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is 'significant', stop; otherwise
  3. update the current model to include the most significant new term; go to 1
- ☐ **Backward elimination**: starting from model with all terms,
  1. if all terms are 'significant', stop; otherwise
  2. update current model by dropping the 'least significant' term; go to 1
- ☐ **Stepwise**: starting from an arbitrary model,
  1. consider three options—add a term, delete a term, swap a term in the model for one not in the model;
  2. if model unchanged, stop; otherwise go to 1
- ☐ 'Significant' might be assessed using F tests, or AIC comparison, or ...

Stepwise methods: Comments

- ☐ Systematic search minimising AIC or similar over all possible models is preferable, but is often infeasible.
- ☐ Original formulation of stepwise used F tests (or even arbitrary numbers!) to assess significance, but this finds spurious models.
- ☐ Nowadays compare AIC for different models at each step—uses AIC (or AIC$_c$) as objective function.
- ☐ Important not to fixate on a specific model, or assume that there is a single 'best' model, but to consider any models whose AIC is within (say) 2 of the minimum.

Example: Nuclear power stations

```
> nuclear
     cost  date t1 t2  cap pr ne ct bw cum.n pt
1  460.05 68.58 14 46  687  0  1  0  0    14  0
2  452.99 67.33 10 73 1065  0  0  1  0     1  0
3  443.22 67.33 10 85 1065  1  0  1  0     1  0
4  652.32 68.00 11 67 1065  0  1  1  0    12  0
5  642.23 68.00 11 78 1065  1  1  1  0    12  0
6  345.39 67.92 13 51  514  0  1  1  0     3  0
7  272.37 68.17 12 50  822  0  0  0  0     5  0
8  317.21 68.42 14 59  457  0  0  0  0     1  0
9  457.12 68.42 15 55  822  1  0  0  0     5  0
10 690.19 68.33 12 71  792  0  1  1  1     2  0
...
32 270.71 67.83  7 80  886  1  0  0  1    11  1
```


Example: Nuclear power stations

	Full model		Backward		Forward	
	Est (SE)	t	Est (SE)	t	Est (SE)	t
Constant	-14.24 (4.229)	-3.37	-13.26 (3.140)	-4.22	-7.627 (2.875)	-2.66
date	0.209 (0.065)	3.21	0.212 (0.043)	4.91	0.136 (0.040)	3.38
log(T1)	0.092 (0.244)	0.38				
log(T2)	0.290 (0.273)	1.05				
log(cap)	0.694 (0.136)	5.10	0.723 (0.119)	6.09	0.671 (0.141)	4.75
PR	-0.092 (0.077)	-1.20				
NE	0.258 (0.077)	3.35	0.249 (0.074)	3.36		
CT	0.120 (0.066)	1.82	0.140 (0.060)	2.32		
BW	0.033 (0.101)	0.33				
log(N)	-0.080 (0.046)	-1.74	-0.088 (0.042)	-2.11		
PT	-0.224 (0.123)	-1.83	-0.226 (0.114)	-1.99	-0.490 (0.103)	-4.77
s (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Selection effects

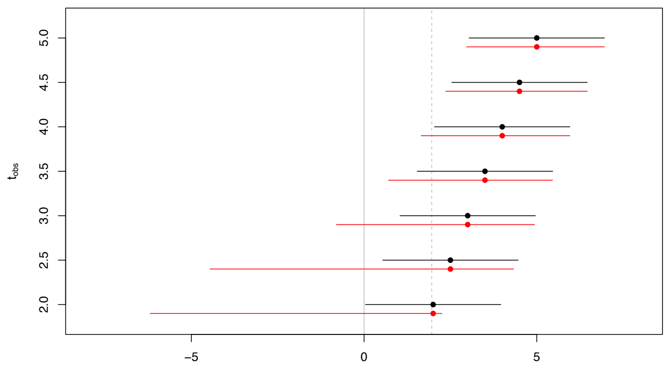
- Contrast
  - **exploratory analysis**, where we study data with no strong prior hypotheses, aiming to find something 'interesting' for future study, and
  - **confirmatory analysis**, where we specify an analysis protocol (hypotheses/tests/...) in advance and stick to it.
- Most statistical procedures assume we are doing the second, but there can be a strong temptation to cheat and treat an exploratory analysis as confirmatory.
- In 'the garden of forking paths' we make a series of choices (which response? transformation? which explanatory variables? ...) but do not then allow for them.
- This leads to non-reproducible results, 'false discoveries', bad science ...
- If we compute a confidence interval $\mathcal{I}$ for θ following a sequence of choices summarised in a selection event S that is *based on the data*, and compute

$$\mathrm{P}(\theta \in \mathcal{I}) \quad \text{when we should compute} \quad \mathrm{P}(\theta \in \mathcal{I} \mid S),$$

we are effectively pretending that S did not exist.

Simple example

Example 15 *Suppose $T \sim \mathcal{N}(\theta, 1)$ and we perform a two-sided test of $H_0 : \theta = 0$ at level $\alpha = 5\%$ and then construct a 95% confidence interval around the observed t_{obs} if we reject H_0. Compare the resulting confidence intervals when we do and do not allow for selection.*



95% confidence intervals for θ without (black) and with (red) allowance for selection on event $S = \{T > z_{0.975}\}$.

Note to Example 15

- ☐ Recall the basis of confidence intervals for θ based on an estimator T satisfying $T \sim \mathcal{N}(\theta, 1)$. We use the fact that $T - \theta \sim \mathcal{N}(0, 1)$ to argue that

$$\mathrm{P}(T \le t_{\mathrm{obs}}) = \mathrm{P}(T - \theta \le t_{\mathrm{obs}} - \theta) = \Phi(t_{\mathrm{obs}} - \theta)$$

  and then set this equal to $\alpha, 1 - \alpha$ to obtain the $(1 - 2\alpha)$ confidence interval $(t_{\mathrm{obs}} - z_{1-\alpha}, t_{\mathrm{obs}} - z_{\alpha})$, which reduces to the 95% confidence interval $t_{\mathrm{obs}} \pm 1.96$ when $\alpha = 0.025$.

- ☐ If we condition on the selection event that $|T| > z_{1-\beta}$ and, if this event occurs, compute the 95% confidence interval for θ, we are effectively using the conditional distribution

$$\begin{aligned} \mathrm{P}(T \le t_{\mathrm{obs}} \mid T > z_{1-\beta}) &= \mathrm{P}(T - \theta \le t_{\mathrm{obs}} - \theta \mid T - \theta > z_{1-\beta} - \theta) \\ &= \frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} \end{aligned}$$

  and the $(1 - 2\alpha)$ interval for θ has as endpoints the solutions of the equations

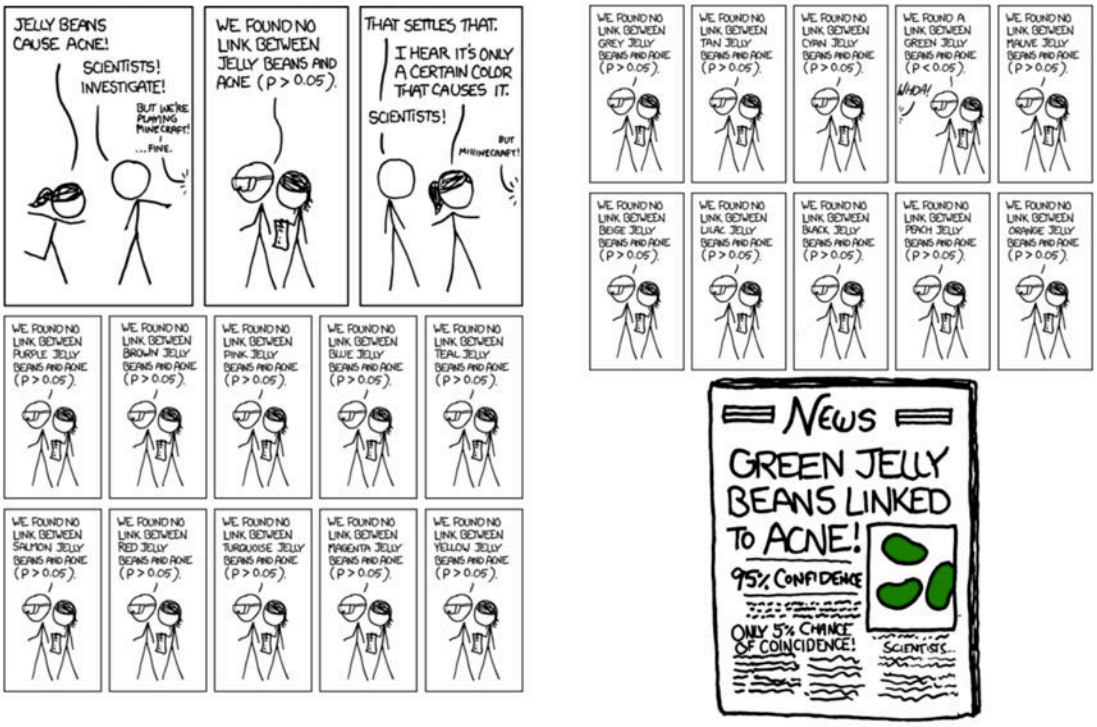
$$\frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} = \alpha, 1 - \alpha.$$

- ☐ If we set $\beta = 0.025$ and $\alpha = 0.025$, then we get the limits shown in the graph, which shows that even having $t_{\mathrm{obs}} = 3$ still leads to a 95% CI that contains 0 when we allow for selection. Hence making allowance for selection can radically change inferences, especially when H_0 is only just rejected.

Implications

- ☐ Need to be aware of possibility of selection effects and to read the literature critically.
- ☐ Must be clear if a study is exploratory or confirmatory:
  - if confirmatory, need to clarify protocol for inference **beforehand**;
  - if exploratory, need to avoid (any?) conclusions that might be due to ‘forking paths’.
- ☐ Active area of research, likely to change in next few years.

Selective reporting of results



M-estimation

- The least squares estimates are linear in y and therefore very sensitive to outliers.
- When $y_i \mapsto y_i + c$,

$$\widehat{\beta} = \sum_{j=1}^{n} (X^{\mathrm{T}}X)^{-1} x_j y_j \mapsto \sum_{j=1}^{n} (X^{\mathrm{T}}X)^{-1} x_j y_j + (X^{\mathrm{T}}X)^{-1} x_i c = \widehat{\beta} + (X^{\mathrm{T}}X)^{-1} x_i c,$$

  which could be arbitrarily far from $\widehat{\beta}$.
- Try and fix this by replacing

$$\min_{\beta} \sum_{j=1}^{n} (y_j - x_j^{\mathrm{T}}\beta)^2 \qquad \text{by} \qquad \min_{\beta} \sum_{j=1}^{n} \rho\left\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\right\},$$

  for function $\rho(\cdot)$ that will give a more robust **M**(aximum likelihood-like)**-estimator**, or equivalently solving the $p \times 1$ system of **estimating equations**

$$\frac{1}{\sigma} \sum_{j=1}^{n} x_j \rho'\left\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\right\} = X^{\mathrm{T}}\rho' = 0$$

  say, where $\rho'_{n\times 1}$ has jth element $\mathrm{d}\rho(u)/\mathrm{d}u$ for $u = (y_j - x_j^{\mathrm{T}}\beta)/\sigma$.

Choice of ρ

- Choose $\rho(u)$ to have desirable properties, e.g., to downweight outliers:

$$\begin{aligned}
\rho(u) &= u^2/2 \quad \text{(normal errors)},\\
\rho(u) &= |u| \quad \text{(Laplace errors)},\\
\rho(u) &= \nu \log(1 + u^2/\nu)/2 \quad (t_\nu \text{ errors}),\\
\rho(u) &= \begin{cases} u^2/2, & |u| < c,\\ c(2|u| - c)/2, & \text{otherwise}, \end{cases} \quad \text{(Huber function)}.
\end{aligned}$$

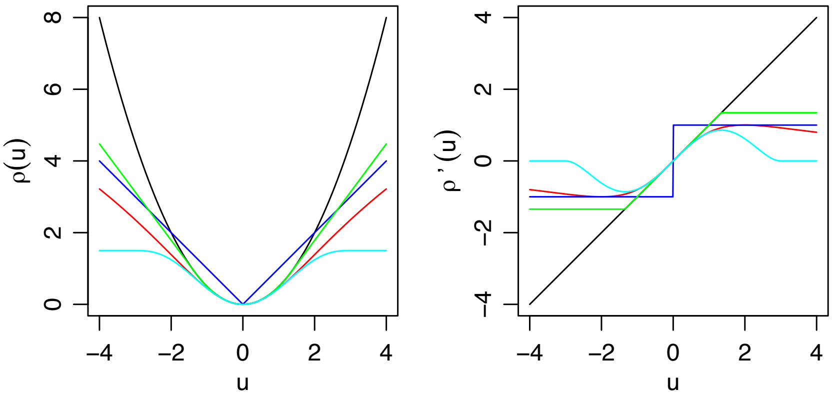
- The function $\rho'(u)$ is also called the **influence function** of the estimator, as its value determines what influence an observation at u has on the estimator:
  - Huber $\rho'(u)$ is bounded,
  - t_ν function is bounded and **redescending**, as $\lim_{u\to\pm\infty} \rho'(u) = 0$;
  - Tukey's **biweight**

$$\rho'(u) = u\left\{1 - (u/c)^2\right\}^2 I(|u| < c),$$

    which gives $\rho'(u) = 0$ when $|u| > c$, is also redescending, giving no weight to observations outside $\pm c$.

ρ and ρ'

Functions ρ and ρ' for least squares (black), t_5 (red), Laplace (blue), Huber (green) and biweight (cyan) estimators.



Estimation

- We need to solve

$$X^{\mathrm{T}}\rho' = 0,$$

where ρ' has jth element

$$\sigma^{-1}\rho'\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\} \propto \frac{\rho'\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\}}{y_j - x_j^{\mathrm{T}}\beta} \times (y_j - x_j^{\mathrm{T}}\beta) = w_j(\beta,\sigma)(y_j - x_j^{\mathrm{T}}\beta),$$

say, so we write the estimating equation as

$$X^{\mathrm{T}}W(y - X\beta) = 0,$$

with $W = \mathrm{diag}\{w_1(\beta,\sigma),\ldots,w_n(\beta,\sigma)\}$.

- We use **iterative weighted least squares**: choose some initial $\tilde\beta$ and σ, then iterate to convergence the steps
  - compute W using the current $\tilde\beta$,
  - compute the weighted least squares estimate,

$$\tilde\beta = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy.$$

- Estimate σ using median absolute deviation of residuals $y_j - x_j^{\mathrm{T}}\tilde\beta$ at each iteration, or similar robust scale estimate.

M-estimator variance

- Estimator $\tilde{\beta}$ is solution to $p \times 1$ system of equations

$$g(y;\beta) = X^{\mathrm{T}}\rho' = 0.$$

- Can show that if the estimating function g is **unbiased**, i.e.

$$\mathrm{E}\left\{g(Y;\beta);\beta\right\} = 0, \quad \text{for any } \beta,$$

then under mild regularity conditions

$$\tilde{\beta} \,\dot{\sim}\, \mathcal{N}_p\left(\beta, \mathrm{E}\left\{-\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}}\right\}^{-1} \mathrm{var}\left\{g(Y;\beta)\right\} \mathrm{E}\left\{-\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}}\right\}^{-1}\right).$$

This is another **sandwich** variance matrix, with

$$\mathrm{E}\left\{-\frac{\partial g(Y;\beta)}{\partial \beta^{\mathrm{T}}}\right\} = X^{\mathrm{T}}W_1X, \quad \mathrm{var}\left\{g(Y;\beta)\right\} = X^{\mathrm{T}}W_2X,$$

so if $W_1 = A(\sigma)I_n$, $W_2 = \sigma^2 B(\sigma)I_n$, then

$$\mathrm{var}(\tilde{\beta}) \doteq \sigma^2(X^{\mathrm{T}}X)^{-1} \times B(\sigma)/A(\sigma)^2.$$

Note: Sandwich matrix I

- The $p \times 1$ estimating function is

$$g(y;\beta) = \sum_{j=1}^{n} x_j \rho'\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right),$$

and unbiasedness implies that if the individual densities are $\sigma^{-1} f\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\}$, then

$$0 = \mathrm{E}\left\{g(y;\beta)\right\} = \sum_{j=1}^{n} x_j \int \rho'\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\sigma^{-1} f\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right) \mathrm{d}y_j = X^{\mathrm{T}} a_{n\times 1},$$

say, where a_j is the jth integral above, and setting $u = (y_j - x_j^{\mathrm{T}}\beta)/\sigma$ shows that all the a_j equal

$$\int \rho'(u) f(u)\,\mathrm{d}u = 0; \quad (5)$$

this is true by symmetry if the error distribution and ρ' are symmetric around the origin. Now

$$\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}} = -\frac{1}{\sigma}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \rho''\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right),$$

whose expectation is (using the same transformation)

$$\begin{aligned}\mathrm{E}\left\{\frac{\partial g(y;\beta)}{\partial \beta^{\mathrm{T}}}\right\} &= -\frac{1}{\sigma}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \mathrm{E}\left\{\rho''\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\} \\ &= -\frac{1}{\sigma}\sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \int \rho''(u) f(u)\,\mathrm{d}u = -\frac{1}{\sigma} X^{\mathrm{T}} X A(\sigma),\end{aligned}$$

say.

- The components of these sums are independent, so

$$\mathrm{var}\left\{g(Y;\beta)\right\} = \mathrm{var}\left\{\sum_{j=1}^{n} x_j \rho'\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right),\right\} = \sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \mathrm{var}\left\{\rho'\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\},$$

where the substitution $u = (y_j - x_j^{\mathrm{T}}\beta)/\sigma$ and (5) show that the variance term can be written as

$$\mathrm{var}\left\{\rho'\left(\frac{Y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)\right\} = \int \rho'(u)^2 f(u)\,\mathrm{d}u = B(\sigma).$$

- The sandwich variance formula is therefore

$$\left\{-\frac{1}{\sigma} X^{\mathrm{T}} X A(\sigma)\right\}^{-1} X^{\mathrm{T}} X B(\sigma) \left\{-\frac{1}{\sigma} X^{\mathrm{T}} X A(\sigma)\right\}^{-1} = (X^{\mathrm{T}} X)^{-1} \times \frac{\sigma^2 B(\sigma)}{A(\sigma)^2}.$$

The variance of the LSE is $\mathrm{var}(Y_j)(X^{\mathrm{T}}X)^{-1}$, so the asymptotic relative efficiency of the M-estimator based on ρ and the LSE is

$$\frac{\mathrm{var}(Y_j)}{\sigma^2} \times \frac{A(\sigma)^2}{B(\sigma)}.$$

Note: Sandwich matrix II

- As a check on this, note that for the normal distribution $\rho'(u) = u$, $f(u) = (2\pi)^{-1}e^{-u^2/2}$, so $A(\sigma) = B(\sigma) = 1$, which gives ARE of 1. If we take $\rho'(u) = \mathrm{sign}(u)$ with the normal density, we have $B(\sigma) = 1$, $A(\sigma) = -2/(2\pi)^{1/2}$, so the sandwich variance formula gives $\sigma^2(X^{\mathrm{T}}X)^{-1}\pi/2$. So using the ρ-function corresponding to the Laplace distribution when the data are in fact normally distributed leads to an estimator which is $\pi/2 \approx 1.57$ times more variable than would be the case if the appropriate ρ-function were used.
- If we take the ρ-function $\rho'(u) = u$ corresponding to the normal density, and the errors are in fact Laplace, $g(u) = (1/2)e^{-|u|}$, we have

$$A(\sigma) = \int (-1)f(u)\,\mathrm{d}u = 1, \quad B(\sigma) = \int u^2 f(u)\,\mathrm{d}u = 2$$

  and the asymptotic relative efficiency is $1/2$.

Efficiency

- Efficiency of M-estimators of β relative to LSEs of β is

$$\frac{\mathrm{var}(Y_j)}{\sigma^2} \times \frac{A(\sigma)^2}{B(\sigma)};$$

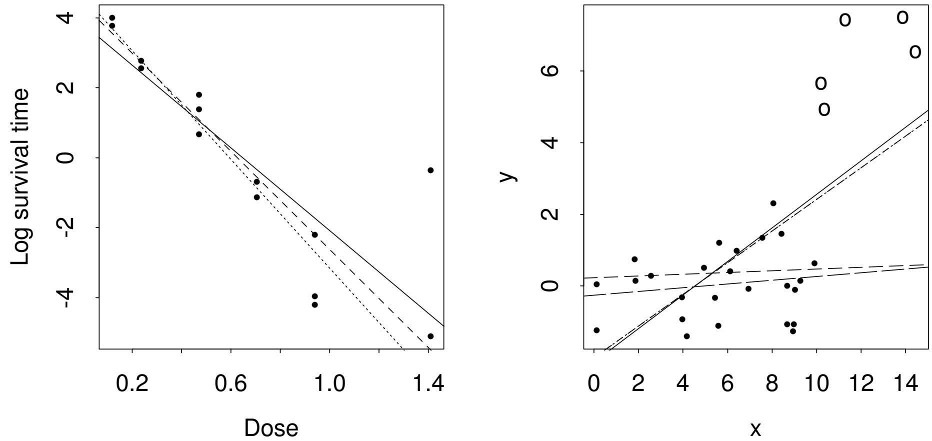
  for example, the Huber estimator is 95% efficient if $c = 1.345$.
- In practice need to balance robustness and efficiency, increasing the latter by increasing c.
- High numbers of outliers can wreck M-estimators.
- Highly robust **least trimmed squares** estimators obtained by minimising

$$\sum_{j=1}^{q} (y_j - x_j^{\mathrm{T}}\beta)^2_{(j)},$$

  where $q = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$.

Example: Survival data

Left: log survival proportions for rats given doses of radiation, with lines fitted by least squares with (solid) and without (dots) the outlier, and a Huber fit for the entire data (dashes). Right: simulated data with a batch of outliers (circles), and fits by least squares to all data (solid), least squares to good data only (large dash), Huber (dot-dash), biweight (dashes), and least trimmed squares (medium dash). The Huber and biweight fits are the same to plotting accuracy.



Simulation (right-hand panel on slide 83)

Table 1: Bias (standard deviation) of estimators of slope in sample of 25 good data and k outliers, estimated from 200 replications.

k	Least squares		M-estimation		Least trimmed squares
	No outliers	With outliers	Huber	Biweight	
1	0.00 (0.07)	0.17 (0.06)	0.07 (0.07)	0.01 (0.07)	-0.01 (0.13)
2	0.00 (0.07)	0.26 (0.06)	0.13 (0.07)	0.02 (0.09)	0.01 (0.14)
5	0.00 (0.07)	0.41 (0.05)	0.38 (0.06)	0.19 (0.19)	0.01 (0.14)
10	0.00 (0.06)	0.48 (0.04)	0.48 (0.04)	0.46 (0.12)	0.05 (0.20)

Good strategy is initial fit using least trimmed squares, then robust fit using this as starting point.

Quantile regression

- The Laplace distribution has

$$\rho(u) = uI(u \geq 0) - uI(u < 0),$$

and for continuous Y, the solution to $\mathrm{E}\{\rho'(Y - \theta)\} = 0$ is the median of Y. Hence

$$\operatorname{argmin} \sum_{j=1}^{n} \rho(y_j - x_j^{\mathrm{T}}\beta)$$

estimates the median of y as a linear function of $X\beta$.

- **Quantile regression** takes $\tau \in (0, 1)$ and uses the **check function**

$$\rho_\tau(u) = \tau uI(u \geq 0) - (1 - \tau)uI(u < 0);$$

then

$$\tilde{\beta}_\tau = \operatorname{argmin} \sum_{j=1}^{n} \rho_\tau(y_j - x_j^{\mathrm{T}}\beta)$$

estimates the τ quantile of y as a linear function of $X\beta$.

- For numerical purposes it may be better to smooth the bottom ρ.
- Note that $\rho_\tau''(u) = 0$, so it's better to bootstrap to find $\mathrm{var}(\tilde{\beta}_\tau)$.

Expectile regression

- Quantile regression can be used to estimate value-at-risk in finance settings, but it has the drawback of just counting how many residuals are above/below the quantile.
- **Expectile regression** extends the LSE in the same way, taking

$$\rho_\tau(y - \theta) = \eta_\tau(y - \theta) - \eta_\tau(y), \quad \eta_\tau(u) = |I(u \leq 0) - \tau|u^2,$$

so $\tau = 1/2$ gives the LSE, while taking $\tau > 1/2$ leads to a more general form of LSE, with good properties for risk estimation in finance applications (coherent elicitable risk measure).

Tall and wide regressions

- So far we have supposed that we have a **tall regression**:
  - the number of units n exceeds the number of variables p,
  - the design matrix X has rank p.
- In many 'modern' settings we instead have a **wide regression**:
  - n and p are comparable, $p > n$, maybe even $p \gg n$;
  - in genomics, for example (typically) $n = O(10^2, 10^3)$, $p = O(10^5, 10^6)$;
  - hence rank$(X) = \min(n, p) = n$.
- Even tall X may be 'almost singular', making β 'almost inestimable'.
- Solutions:
  - subset selection (drop certain columns of X);
  - regularisation (often with prediction in mind);
  - seek different good explanations of response variation, not single model.

Semi-descriptive analysis

- With $p > n$, perhaps $p \gg n$, X is rank-deficient and (perhaps) many β give $X\beta = y$.
- To find important variables we include intrinsic variables (gender, ...) in all models, and then
  - choose some k (preferably ≤ 15) such that $k < n$ and suppose that $p < k^a$ (let $a = 3$ for easy visualisation);
  - assign each variable to a cell of a hyper-cube with coordinates $\{1, \dots, k\}^a$;
  - fit a linear model containing each set of k variables corresponding to the ak^{a-1} rows, columns, ... of the cube, so each variable appears in a distinct models;
  - for each such model, retain the two variables that are most significant.
- Iterate the above procedure, retaining only the significant variables at each stage, aiming for a final set of 10–20 variables, for which a careful analysis is performed, perhaps leading to several different good explanations of the response variation.
- Some cells of the hyper-cube may be empty, and important variables might be assigned to several cells.
- The above design is a form of **balanced incomplete block design (BIBD)** (with k^a treatments and ak^{a-1} blocks).
- See Cox and Battey (2017, PNAS)

Singular value decomposition

- ☐ Write

$$X_{n\times p} = U_{n\times n} D_{n\times p} V^{\mathrm{T}}_{p\times p}$$

  where

  - $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_p)$ are orthogonal and D is $n \times p$ diagonal with diagonal entries (**singular values**) $d_1 \geq \cdots \geq d_m \geq 0$, where $m = \min(n, p)$,
  - if one or more $d_j = 0$, then X is singular, and
  - the u_j and v_r respectively span the column and row spaces of X.

- ☐ If the columns of X are centered (i.e., sum to zero), the spectral (eigen) decomposition of the sample covariance matrix $n^{-1}X^{\mathrm{T}}X$ is

$$n^{-1} V D^{\mathrm{T}} D V^{\mathrm{T}},$$

  and the v_r are the **principal components (Karhunen–Loève directions)** of X, i.e.,

$$X v_1 = u_1 d_1 = z_1, X v_2 = u_2 d_2 = z_2, \ldots,$$

  say, have the largest, second largest, ...variances of the normalised linear combinations of the columns of X, with

$$n^{-1} z_r^{\mathrm{T}} z_r = n^{-1}(X v_r)^{\mathrm{T}}(X v_r) = d_r^2/n.$$

Collinearity

- ☐ Columns of X **collinear** if there exists a non-zero vector $v_{p\times 1}$ such that $Xv = 0$, i.e., $\mathrm{rank}(X) < p$.
- ☐ Then $X^{\mathrm{T}}X$ has no unique inverse, and $\widehat{\beta}$ is not unique.
- ☐ Similar problems if $Xv \doteq 0$: consider distribution of the (squared) distance of $\widehat{\beta}$ from its target β, i.e.,

$$Q = \|\widehat{\beta} - \beta\|^2 = (\widehat{\beta} - \beta)^{\mathrm{T}}(\widehat{\beta} - \beta),$$

  and find (in normal model) that

$$\mathrm{E}(Q) = \sigma^2 \sum_{r=1}^{p} d_r^{-2}, \quad \mathrm{var}(Q) = 2\sigma^4 \sum_{r=1}^{p} d_r^{-4}.$$

- ☐ Collinearity often measured using **condition number** $(d_p/d_1)^{1/2}$, but its statistical meaning is unclear.
- ☐ Simplest solution: drop columns from X. But which?

Note on collinearity

- When $p < n$ and $m = p$ we have $y \sim (X\beta, \sigma^2 I_n) \sim (UD\gamma, \sigma^2 I_n)$, where $\gamma = V^{\mathrm{T}}\beta$, and as γ is just an orthogonal transformation of β and (in this case) $\widehat{\gamma} = V^{\mathrm{T}}\widehat{\beta}$, we have

$$(\widehat{\beta} - \beta)^{\mathrm{T}}(\widehat{\beta} - \beta) = (\widehat{\gamma} - \gamma)^{\mathrm{T}}(\widehat{\gamma} - \gamma).$$

- Now having $y \sim (UD\gamma, \sigma^2 I_n)$ implies that

$$\widehat{\gamma} = \{(UD)^{\mathrm{T}}UD\}^{-1}(UD)^{\mathrm{T}}y = (D^{\mathrm{T}}D)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}y = \mathrm{diag}(d_1^{-1}, \ldots, d_p^{-1}, 0, \ldots, 0)U^{\mathrm{T}}y,$$

so

$$\mathrm{var}(\widehat{\gamma}) = \sigma^2 \mathrm{diag}(d_1^{-2}, \ldots, d_p^{-2}).$$

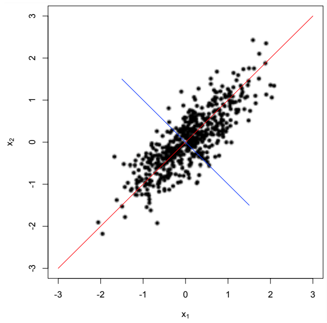
Thus if $d_p \approx 0$, there is at least one direction in which γ, i.e., $v^{\mathrm{T}}\beta$ for some $v_{p\times 1}$, is extremely poorly determined.

- Under the normal model the $\widehat{\gamma}_r$ are independent $\mathcal{N}(\gamma_r, \sigma^2/d_r^2)$, so $\widehat{\gamma}_r - \gamma_r \stackrel{\mathrm{D}}{=} \sigma Z_r/d_r$, where $Z_1, \ldots, Z_p \stackrel{\mathrm{iid}}{\sim} \mathcal{N}(0, 1)$, giving

$$Q = (\widehat{\gamma} - \gamma)^{\mathrm{T}}(\widehat{\gamma} - \gamma) \stackrel{\mathrm{D}}{=} \sum_{r=1}^{p} \sigma^2 Z_r^2/d_r^2,$$

and as $\mathrm{E}(Z_r^2) = 1$ and $\mathrm{var}(Z_r^2) = 2$ we get the desired results.

Principal component regression



- Assume columns of X have been centered, then
  - regress y on $u_1, \ldots, u_{m'}$, then
  - choose $m' \leq m$ to trade off model dimension m' and good fit.
- Used to reduce variance, to deal with collinearity, for dimension reduction, and regularisation, but generally hard to interpret results in terms of original columns of X.
- Assumes that high variation in X corresponds to high variation in y: is this true?
- Alternative: $\max_{a_1} \mathrm{corr}(y, Xa_1)$, then $\max_{a_2} \mathrm{corr}(y, Xa_2)$ subject to $\|a_r\| = 1$, $a_1^{\mathrm{T}}a_2 = 0$, etc.

Regularisation

- ☐ Subset selection/PC regression take only a subset of (transformed) columns of X, but discreteness (retain/delete) of process increases variance of predictions.
- ☐ Regularisation via shrinkage aims to make this smoother, as more continuity should decrease the overall variance.
- ☐ Typically we first center both y and X so that the intercept is zero, i.e., map

$$y \mapsto (I - H)y, \quad X \mapsto (I_n - H)X, \quad \text{with} \quad H = 1_n(1_n^{\mathrm{T}}1_n)^{-1}1_n^{\mathrm{T}},$$

  sometimes we also rescale to give the columns of X unit variance.
- ☐ Then we minimise the sum of squares subject to a penalty on β, taking

$$\widehat{\beta}_\lambda = \mathrm{argmin}_\beta \quad \|y - X\beta\|_2^2 + \lambda p(\beta), \quad \lambda > 0,$$

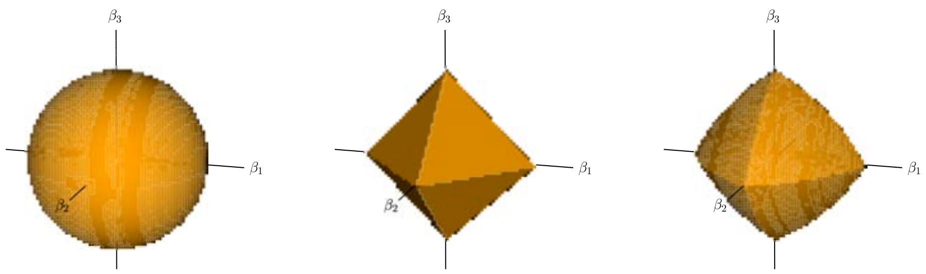
  where among many possibilities,
  - $p(\beta) = \|\beta\|_2^2 = \sum_{r=1}^p \beta_r^2$ gives **ridge regression** (aka Tikhonov regularisation);
  - $p(\beta) = \|\beta\|_1 = \sum_{r=1}^p |\beta_r|$ gives the **lasso** (aka L_1 regularisation);
  - $p(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$ for $0 \le \alpha \le 1$ gives the **elastic net**;
  - $p(\beta) = \sum_{g=1}^G p_g^{1/2}\|\beta_g\|_2$, with β_g being $p_g \times 1$ sub-vectors of β, gives the **grouped lasso**, which penalises factors with parameters β_g.

Bound form

- ☐ Equivalently we can take the **bound form** of the minimisation problem, i.e.,

$$\text{minimise}_\beta \quad \|y - X\beta\|_2^2 \quad \text{subject to} \quad p(\beta) \le t,$$

  for some $t \ge 0$, where setting $t = \infty$ just gives the least squares estimates.
- ☐ Below: constraint balls for ridge (left), lasso (centre) and elastic-net (right) regularisation. The sharp corners of the last two allow for variable selection as well as shrinkage.



Ridge regression

- Writing the criterion in matrix form gives

$$\widehat{\beta}_\lambda = \mathrm{argmin}_\beta \quad (y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\beta^{\mathrm{T}}\beta,$$

  and thus

$$\begin{aligned} \widehat{\beta}_\lambda &= (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}y \\ &= V(D^{\mathrm{T}}D + \lambda I_p)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}y \\ &= \sum_{d_j>0} v_j \times \frac{d_j}{d_j^2 + \lambda} u_j^{\mathrm{T}} y, \end{aligned}$$

  i.e., a linear combination of the v_j with coefficients $u_j^{\mathrm{T}} y d_j/(d_j^2 + \lambda)$ that are more shrunk towards zero when d_j is small.
- The fitted value is a linear combination of the principal components of X,

$$\widehat{y}_\lambda = X\widehat{\beta}_\lambda = H_\lambda y = \sum_{d_j>0} u_j \times \frac{1}{1 + \lambda/d_j^2} u_j^{\mathrm{T}} y,$$

  with increasing shrinkage as j increases and λ/d_j^2 increases (sketch).
- Hence the SVD of X gives the ridge solution for all $\lambda \geq 0$.

Bias-variance tradeoff

Lemma 16 *Under the second-order assumptions $y \sim (X\beta, \sigma^2 I_n)$, with $n > p$ and rank$(X) = p$, $\widehat{\beta}_\lambda$ has bias vector and variance matrix*

$$-\sum_{r=1}^{p} v_r \times \frac{1}{1 + d_r^2/\lambda} v_r^{\mathrm{T}}\beta, \quad \sigma^2 \sum_{r=1}^{p} \frac{d_r^2}{(d_r^2 + \lambda)^2} v_r v_r^{\mathrm{T}}.$$

- As $\lambda \to \infty$, the bias increases to β and the variance decreases to zero.
- The interpretation of λ is unclear, but by analogy with the usual linear model we define the degrees of freedom to be

$$\mathrm{tr}(H_\lambda) = \mathrm{tr}\{X(X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}\} = \sum_{r=1}^{p} \frac{d_r^2}{d_r^2 + \lambda},$$

  which
  - starts at p provided $d_p > 0$, corresponding to a standard least squares fit with $\lambda = 0$, and
  - is monotone decreasing in λ, and hence gives another more interpretable measure of regularity.

Note to Lemma 16

- ☐ The objective function for ridge regression is

$$y^{\mathrm T}y - 2y^{\mathrm T}X\beta + \beta^{\mathrm T}(X^{\mathrm T}X + \lambda I_p)\beta,$$

and differentiation with respect to β gives first and second derivatives

$$-2yX^{\mathrm T}y + 2(X^{\mathrm T}X + \lambda I_p)\beta, \quad X^{\mathrm T}X + \lambda I_p.$$

The second derivative matrix is positive definite for any $\lambda > 0$, and setting the first to 0 gives

$$\widehat{\beta}_\lambda = (X^{\mathrm T}X + \lambda I_p)^{-1}X^{\mathrm T}y.$$

- ☐ Setting $X = UDV^{\mathrm T}$ gives $X^{\mathrm T}y = VD^{\mathrm T}U^{\mathrm T}y = \sum_{d_j>0} v_j d_j u_j^{\mathrm T}y$ and

$$(X^{\mathrm T}X + \lambda I_p)^{-1} = (VD^{\mathrm T}DV^{\mathrm T} + \lambda I_p)^{-1} = \{V(D^{\mathrm T}D + \lambda I_p)V^{\mathrm T}\}^{-1} = VS_\lambda V^{\mathrm T},$$

where $S_\lambda = \mathrm{diag}(d_1^2 + \lambda, \ldots, d_r^2 + \lambda)^{-1}$ exists because all its elements are positive. Hence

$$\widehat{\beta}_\lambda = (X^{\mathrm T}X + \lambda I_p)^{-1}X^{\mathrm T}y = VS_\lambda V^{\mathrm T}(VD^{\mathrm T}U^{\mathrm T})y = \sum_{d_j>0} \frac{d_j}{d_j^2 + \lambda} u_j^{\mathrm T}y \times v_j.$$

Note II to Lemma 16

- ☐ Under the second-order assumptions

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_\lambda) &= VS_\lambda V^{\mathrm T}VD^{\mathrm T}U^{\mathrm T}UDV^{\mathrm T}\beta = \sum_{r=1}^{p} \frac{d_r^2}{d_r^2 + \lambda} v_r^{\mathrm T}\beta \times v_r, \\
\mathrm{var}(\widehat{\beta}_\lambda) &= VS_\lambda D^{\mathrm T}U^{\mathrm T}\mathrm{cov}(y)\{VS_\lambda D^{\mathrm T}U^{\mathrm T}\}^{\mathrm T} = \sigma^2 V \mathrm{diag}\left\{\frac{d_1^2}{(d_1^2+\lambda)^2}, \ldots, \frac{d_p^2}{(d_p^2+\lambda)^2}\right\} V^{\mathrm T}.
\end{aligned}$$

The latter is just the given formula for the variance, and the bias is

$$\mathrm{E}(\widehat{\beta}_\lambda) - \beta = \sum_{r=1}^{p} \frac{d_r^2}{d_r^2 + \lambda} v_r^{\mathrm T}\beta \times v_r - \sum_{r=1}^{p} v_r v_r^{\mathrm T}\beta = -\sum_{r=1}^{p} \frac{\lambda}{d_r^2 + \lambda} v_r^{\mathrm T}\beta \times v_r$$

- ☐ Finally,

$$\mathrm{tr}(H_\lambda) = \mathrm{tr}\{(UDV^{\mathrm T})VS_\lambda V^{\mathrm T}(UDV^{\mathrm T})^{\mathrm T}\} = \mathrm{tr}\{U^{\mathrm T}UDS_\lambda D^{\mathrm T}\} = \sum_{r=1}^{p} \frac{d_r^2}{d_r^2 + \lambda},$$

as required.

Choice of λ

☐ Can aim to choose λ for best prediction at an individual x_+, minimising

$$\mathrm{MSE}(x_+, \lambda) = \mathrm{E}\{(x_+^{\mathrm{T}}\widehat{\beta} - x_+^{\mathrm{T}}\beta)^2\} = \{x_+^{\mathrm{T}}\mathrm{Bias}(\widehat{\beta}_\lambda)\}^2 + x_+^{\mathrm{T}}\mathrm{var}(\widehat{\beta}_\lambda)x_+^{\mathrm{T}},$$

or the generalisation to a matrix X_+.

☐ Usually choose λ to minimise the (generalized) cross-validation sums of squares

$$\mathrm{CV}(\lambda) = \sum_{j=1}^{n}(y_j - \widehat{y}_{\lambda,-j})^2, \quad \mathrm{GCV}(\lambda) = \sum_{j=1}^{n}\frac{(y_j - \widehat{y}_{\lambda,j})^2}{\{1 - \mathrm{tr}(H_\lambda)/n\}^2},$$

where $\widehat{y}_{\lambda,-j}$ is the fitted value for y_j predicted from a ridge fit without case j.

Lemma 17 *For a fit with $\widehat{y} = Hy$ where H has $j\mathrm{th}$ diagonal element h_{jj},*

$$\sum_{j=1}^{n}(y_j - \widehat{y}_{-j,j})^2 = \sum_{j=1}^{n}\frac{(y_j - \widehat{y}_j)^2}{(1 - h_{jj})^2}.$$

Note to Lemma 17

☐ Suppose we leave out (x_j, y_j) and solve

$$\text{minimise}_\beta \quad \sum_{i\neq j}(y_i - x_i^{\mathrm{T}}\beta)^2 + \lambda p(\beta),$$

leading to solution $\widehat{\beta}_{-j}$. Let $y_j^* = \widehat{y}_{-j,j} = x_j^{\mathrm{T}}\widehat{\beta}_{-j}$ be the corresponding fitted value for x_j.

☐ Inserting back the pair (x_j, y_j^*) into the dataset used to compute $\widehat{\beta}_{-j}$ changes nothing, because $(y^* - x_j^{\mathrm{T}}\widehat{\beta}_{-j})^2 = 0$ and $p(\beta)$ does not depend on the data. Now For this new dataset,

$$y_j^* = \sum_{i\neq j} h_{ji}y_i + h_{jj}y_j^* = \sum_i h_{ji}y_i + h_{jj}(y_j^* - y_j) = \widehat{y}_j + h_{jj}(y_j^* - y_j)$$

which implies that

$$y_j - y_j^* = y_j - \widehat{y}_j + h_{jj}(y_j - y_j^*),$$

leading to

$$y_j - y_j^* = y_j - \widehat{y}_{-j,j} = \frac{y_j - \widehat{y}_j}{1 - h_{jj}},$$

and thus to the given formula.

☐ Note that the argument above applies to any linear fit.

Example: Cement data

```
> cement
   x1 x2 x3 x4     y
1   7 26  6 60  78.5
2   1 29 15 52  74.3
3  11 56  8 20 104.3
4  11 31  8 47  87.6
5   7 52  6 33  95.9
6  11 55  9 22 109.2
7   3 71 17  6 102.7
8   1 31 22 44  72.5
9   2 54 18 22  93.1
10 21 47  4 26 115.9
11  1 40 23 34  83.8
12 11 66  9 12 113.3
13 10 68  8 12 109.4
```

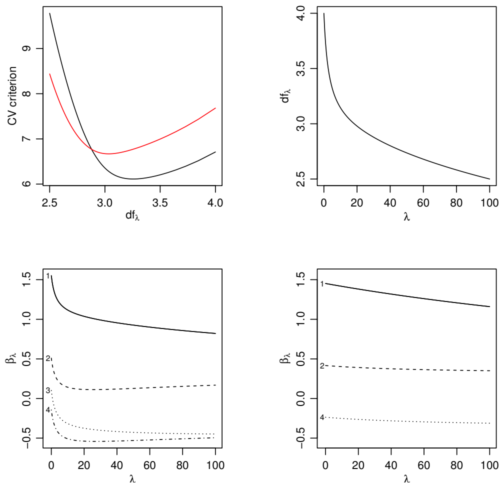

Example: Cement data

	Full model		Reduced model	
Parameter	Estimate	Standard error	Estimate	Standard error
β_0	62.41	70.07	71.64	14.14
β_1	1.55	0.74	1.45	0.12
β_2	0.51	0.72	0.42	0.19
β_3	0.10	0.75		
β_4	–0.14	0.71	–0.24	0.17

- ☐ The next slide shows results for ridge fits for these models.
- ☐ Looks like 3 df is optimal for prediction.
- ☐ The singular values for the centred X matrix are 78.8, 28.5, 12.2, 1.7, and those for the centred and scaled X matrix are 5.18, 4.35, 1.50, 0.14, so it matters which is used.
- ☐ The singular values for the (centred) reduced matrix are 78.8, 19.8 and 9.15.
- ☐ The shrinkage due to increasing λ occurs more slowly for the reduced model

Example: Cement data/Ridge analysis

Top left: CV (black) and GCV (red) as functions of degrees of freedom df_λ. Top right: dependence of df_λ on λ. Bottom left: $\widehat{\beta}_\lambda$ as a function of λ, with all four covariates. Bottom right: $\widehat{\beta}_\lambda$ as a function of λ, with x_1, x_2, and x_4 only.



Comments

- ☐ Be careful with software: any pre-processing of X is not always described.
- ☐ The literature on ridge regression is very large and very dispersed, with many variants and many connections to ML techniques.
- ☐ Bayesian interpretation: $\widehat{\beta}_\lambda$ is the posterior mean/mode for β in the model

$$y \mid \beta \sim \mathcal{N}_n(\beta_0 1_n + X\beta, \sigma^2 I_n), \quad \beta \sim \mathcal{N}_p(0, \sigma^2 I_p/\lambda), \quad \sigma^2, \lambda > 0,$$

  with X centred, an improper uniform prior on β_0 and σ^2 and λ fixed. The latter can be estimated/chosen using **empirical Bayes** or **REML** (later).
- ☐ Similar Bayesian interpretation as posterior mode (not mean) for other penalties.
- ☐ In 'ordinary ridge' the penalisation is the same for all elements of β and shrinkage is towards the origin. More generally, we could minimise

$$\|y - \beta_0 1_n - X\beta\|^2 + \lambda(\beta - \beta')^{\mathrm{T}} W(\beta - \beta'),$$

  corresponding to shrinkage towards β' according to dispersion matrix W, equivalent to taking prior

$$\beta \sim \mathcal{N}_p(\beta', \sigma^2 W^{-1}/\lambda), \quad \sigma^2, \lambda > 0,$$

  above.

Lasso

- **Lasso (least absolute shrinkage and selection operator)** (aka **basis pursuit**) solves

$$\text{minimise}_\beta \quad \|y - X\beta\|_2^2 \quad \text{such that} \quad \|\beta\|_1 \le t,$$

or equivalently finds

$$\tilde\beta_\lambda = \text{argmin}_\beta \quad (y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\|\beta\|_1, \quad \lambda > 0.$$

- If $t > t_0 = \|\widehat\beta\|_1$ then $\lambda = 0$ and $\tilde\beta_\lambda = \widehat\beta$, whereas if $t \approx t_0/2$ then the estimates are shrunk by a factor around 0.5 on average, with

$$\lim_{\lambda\to\infty} \tilde\beta_\lambda = 0, \qquad \lim_{\lambda\to 0} \tilde\beta_\lambda = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \widehat\beta$$

- Orthogonal design matrix $X^{\mathrm{T}}X = I_p$ gives **soft thresholding** function

$$\tilde\beta_{\lambda,r} = g_\lambda(\widehat\beta_r) = \begin{cases} 0, & |\widehat\beta_r| < \lambda, \\ \text{sign}(\widehat\beta_r)(|\widehat\beta_r| - \lambda), & \text{otherwise}, \end{cases} \qquad r = 1, \dots, p,$$

and this happens in general: if the constraints bite, then some of the $\tilde\beta_{\lambda,r}$ are zero.

Note on soft thresholding

- ☐ We have $X^{\mathrm{T}}X = I_p$ and hence $\widehat{\beta} = X^{\mathrm{T}}y$. The Lagrangian

$$L = \tfrac{1}{2}(y^{\mathrm{T}}y - 2y^{\mathrm{T}}X\beta + \beta^{\mathrm{T}}\beta) + \lambda\left(\sum_{r=1}^{p}|\beta_r| - t\right)$$

is a sum of two convex functions and therefore is convex in β. Apart from constants, we can write

$$2L \equiv \sum_{r=1}^{p}(\beta_r^2 - 2\widehat{\beta}_r\beta_r + 2\lambda|\beta_r|) - 2p\lambda t,$$

which is differentiable except at $\beta_r = 0$. This is a sum of p separate functions which can be minimised individually, and in which we replace β_r by the scalar β.

- ☐ In each case, either the minimum is at $\beta = 0$ or elsewhere. Differentiation gives

$$\partial L/\partial\beta = \beta - \widehat{\beta} + \lambda\mathrm{sign}(\beta),$$

and

$$\lim_{\beta\to 0_+}\partial L/\partial\beta = \lambda - \widehat{\beta}, \quad \lim_{\beta\to 0_-}\partial L/\partial\beta = -\lambda - \widehat{\beta}.$$

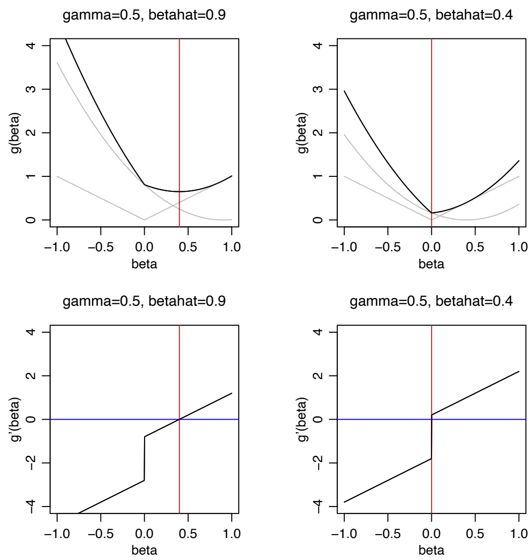
For a minimum at $\beta = 0$ we must have $\lambda - \widehat{\beta} > 0$ and $-\lambda - \widehat{\beta} < 0$, or equivalently $|\widehat{\beta}| < \lambda$. Elsewhere $\partial L/\partial\beta = 0$ gives

$$\tilde{\beta} = \widehat{\beta} - \lambda\mathrm{sign}(\tilde{\beta}),$$

so if $\tilde{\beta} > 0$, then $\tilde{\beta} = \widehat{\beta} - \lambda$, whereas if $\tilde{\beta} < 0$, then $\tilde{\beta} = \widehat{\beta} + \lambda$. This leads to the function $g_\lambda(\cdot)$ given on the slide.

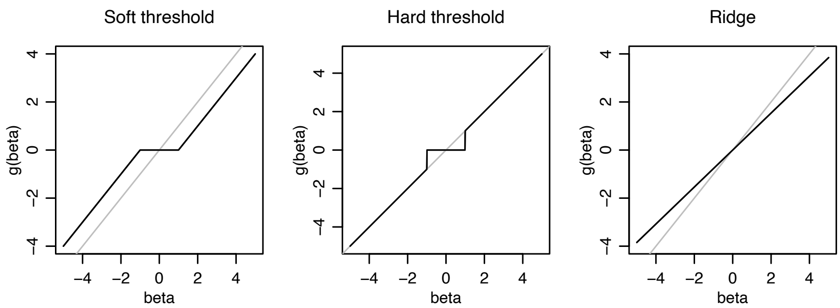
- ☐ The top graphs on slide 103 show grey lines corresponding to $\|\beta|$ and $(\beta - \widehat{\beta})^2$; the black line is their sum. In both cases $\lambda = 0.5$. In the left panel with $\widehat{\beta} = 0.9$ the minimum is at $\beta = 0.4$, and on the right panel with $\widehat{\beta} = 0.4$ the minimum is at $\beta = 0$.
The lower panels show the corresponding derivatives (black) and the value of β (red) at which they equal zero (shown by the blue horizontal line). If $\widehat{\beta}$ is sufficiently far from zero, then the intersection will not be at zero, but otherwise it is.

Soft thresholding



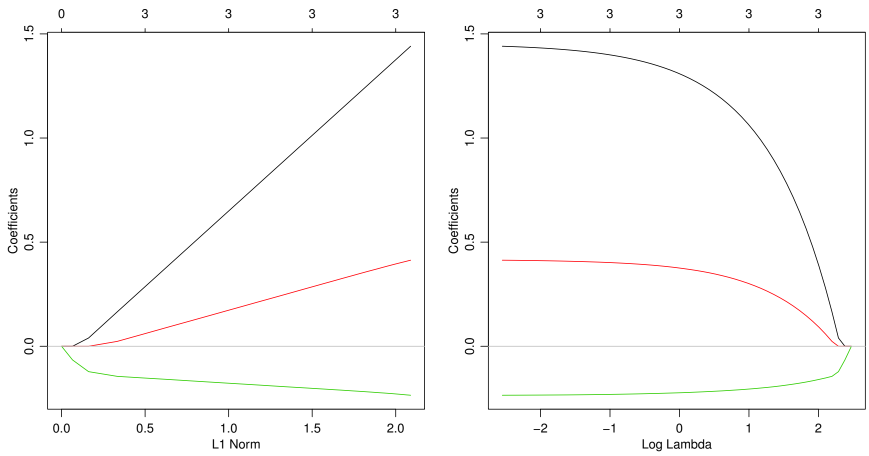
Threshold functions

- Soft threshold of lasso (left), hard threshold of subset selection (middle), no threshold of ridge (right).



Example: cement data

- Estimated coefficients for lasso fit against L_1 norm and λ:



Comments

- **Least angle regression (LAR)** is similar to the lasso, and the LAR algorithm can compute the lasso solution path for all λ in $O(n)$ operations (as with ridge).
- For any regression model we can define the **degrees of freedom** as

$$\sigma^{-2}\sum_{j=1}^{n}\mathrm{cov}(y_j,\widehat{y}_j)=\mathrm{tr}\{\mathrm{cov}(y,\widehat{y})\}/\sigma^2;$$

  this reduces to previous definitions but can be computed in more situations.
- **Theory**: one can ask about the properties of $\tilde{\beta}_\lambda$ in suitable settings (e.g., $n,p\to\infty$ with $p/n\to c>0$). Then under certain conditions one can show that the selection of variables is consistent (i.e., the probability that the variables with $\beta_r\neq 0$ are selected tends to 1), but that the $\tilde{\beta}_\lambda$ are inconsistent (because soft thresholding implies that $|\tilde{\beta}_{\lambda,r}|$ is systematically smaller than $|\beta_r|$).
- Many (many!) variants and related procedures exist to overcome such problems.
- **Computation**: lasso and elastic net penalisations available in `R` package `glmnet` and extend to generalized linear models and more general regressions (later).

Components of Variance

slide 107

Background and motivation

- ☐ All the models so far have involved just one level of randomness, corresponding to 'measurement error' on individual responses.
- ☐ Complex layering of randomness can arise in applications, and then conclusions may depend on how it is dealt with.
- ☐ Two conceptually different set-ups (which may give the same models):
  - – observational/experimental setup generates several layers of randomness;
  - – we find it useful to treat the parameters of some model as drawn from a distribution.

  The first concerns logical properties of the data, whereas the second is a modelling assumption.

Example: Blood pressure

- ☐ Blood pressure data: $P = 25$ patients each made $V = 16$ visits to a clinic, and on each occasion their systolic and diastolic blood pressures were measured twice.
- ☐ Consider just the diastolic pressure. We expect there to be variation
  - – between patients,
  - – between visits within patients, and
  - – between measurements within visits,

  which we could model as

$$y_{pvm} = \mu + b_p + e_{pv} + \varepsilon_{pvm}, \quad p = 1, \ldots, P, v = 1, \ldots, V, m = 1, \ldots, M,$$

  where
  - – μ is the (hypothetical) population mean diastolic blood pressure (DBP),
  - – b_p is the difference between the (hypothetical) patient and population mean DBP,
  - – e_{pv} is the difference between this and the (hypothetical) mean DBP on the vth visit, and
  - – ε_{pvm} is the difference between the mean DBP for the pth patient at the vth visit and the mth measurement on that visit.
- ☐ The existence of some of these hypothetical means may be problematic.

Example: Blood pressure

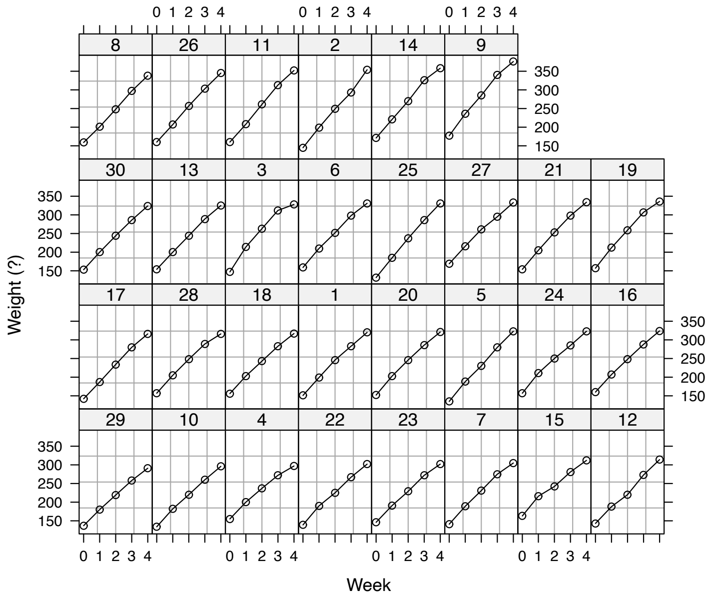
```
patno patient visno dbp1 dbp2 sbp1 sbp2
1307    1         7   95   85  150  130
1307    1         8   85   85  140  140
1307    1         9   90   90  150  150
1307    1        10   80   80  135  135
1307    1        11   80   80  130  125
1307    1        12   85   85  150  155
.
.
1307    1        19   80   80  130  130
1307    1        20   80   80  140  140
1307    1        21   90   85  145  140
1307    1        22   75   75  130  130
1418    2         7  104  106  160  148
1418    2         8   98  104  158  162
.
.
9202   25        21   91   90  142  139
9202   25        22   80   78  162  160
```


Example: Rat growth

Weights (units unknown) of 30 young rats over a five-week period

	Week						Week				
	1	2	3	4	5		1	2	3	4	5
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

Example: Rat growth



Example: Rat growth

☐ Here a natural model is that growth is linear for each rat, i.e.,

$$y_{rw} \mid \alpha_r, \beta_r \overset{\text{ind}}{\sim} \mathcal{N}\{\alpha_r + \beta_r(w-1), \sigma^2\}, \quad r = 1, \ldots, 30, w = 1, \ldots, 5,$$

where ε_{rw} represents measurement variation for each rat and week.

☐ If we are not interested in the particular α_r and β_r, but in population mean values, we might write

$$\alpha_r \overset{\text{iid}}{\sim} \mathcal{N}(\alpha, \sigma_\alpha^2), \quad \beta_r \overset{\text{iid}}{\sim} \mathcal{N}(\beta, \sigma_\beta^2), \quad \text{corr}(\alpha_r, \beta_r) = \rho,$$

and try and estimate the population means $\alpha = \mathrm{E}(\alpha_r)$ and $\beta = \mathrm{E}(\beta_r)$, allowing for

– variation between rats, and
– within rats between measurements.

Fixed and random effects

Chimpanzee	Word									
	1	2	3	4	5	6	7	8	9	10
1	178	60	177	36	225	345	40	2	287	14
2	78	14	80	15	10	115	10	12	129	80
3	99	18	20	25	15	54	25	10	476	55
4	297	20	195	18	24	420	40	15	372	190

- The table shows times (min) for four chimpanzees to learn each of ten words.
- A possible model for log time is

$$y_{cw} \mid \alpha_c, \beta_w \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu + \alpha_c + \beta_w, \sigma^2), \quad c = 1, \ldots, C = 4, w = 1, \ldots, W = 10.$$

- The α_c and/or the β_w would be considered as constant **fixed effects** if we were interested in the relative linguistic abilities of these particular chimps and/or if we planned further tests with these particular words.
- Either (or both) of the α_c and β_w might be considered to be **random effects** if they were thought to be sampled from a larger population whose variation is of interest.

Two distinctions

- We distinguish **fixed** and **random** effects (above).
- We distinguish **nested** and **crossed** effects:
  - in the blood pressure data, replicate measurements at each visit are **nested** within visit, because there is no logical connection between $y_{p,v_1,1}$ and $y_{p,v_2,1}$ (we could permute the final index m within each patient/visit combination without changing the data structure). Likewise if we ignore any possible time effects between visits, we could consider that visits are nested within patients;
  - in the chimp data, the effects are **crossed**, because permuting chimps or words would entail permuting entire rows or columns of the data table: there is a logical connection between y_{c_1w} and y_{c_2w}, and between y_{cw_1} and y_{cw_2};
- In R syntax, with patient and visit number declared as factors, for nested effects we write

```
y ~ patient/visno
```

  read as 'separate effects for visit number within the levels of patient' and for crossed effects with chimp and word declared as factors we write

```
y ~ chimp + word
```

Example

Example 18 (One-way layout) *Consider R units in each of T blocks with the random effects model*

$$\begin{aligned} y_{t,r} \mid b_t &\overset{\text{ind}}{\sim} \mathcal{N}(\mu + b_t, \sigma^2), \quad r = 1, \ldots, R, \\ b_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2), \quad t = 1, \ldots, T. \end{aligned}$$

Find the joint distributions of the responses and of the sums of squares

$$SS_w = \sum_{t=1}^{T}\sum_{r=1}^{R}(y_{tr} - \overline{y}_{t\cdot})^2, \quad SS_b = \sum_{t=1}^{T}\sum_{r=1}^{R}(\overline{y}_{t\cdot} - \overline{y}_{\cdot\cdot})^2 = R\sum_{t=1}^{T}(\overline{y}_{t\cdot} - \overline{y}_{\cdot\cdot})^2$$

within and between blocks. How do you test for $\sigma_b^2 = 0$?

☐ Similar arguments apply in other balanced settings ...

Note to Example 18

☐ Recall that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\sum_{j=1}^{n}(X_j - \overline{X})^2 \sim \sigma^2\chi^2_{n-1}$, and if $V_1, \ldots, V_m \overset{\text{ind}}{\sim} \tau\chi^2_{\nu_1}, \ldots, \tau\chi^2_{\nu_m}$, then $\sum_{j=1}^{m} V_j \sim \tau\chi^2_{\nu_1+\cdots+\nu_m}$.

☐ A convenient way to write the model is

$$y_{t,r} \overset{\text{D}}{=} \mu + b_t + \varepsilon_{t,r}, \quad b_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2) \quad \perp\!\!\!\perp \quad \varepsilon_{t,r} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

which immediately gives the block diagonal covariance structure

$$\text{cov}(y_{t,r}, y_{t',r'}) = \sigma^2 I(t = t', r = r') + \sigma_b^2 I(t = t').$$

Moreover

$$y_{tr} - \overline{y}_{t\cdot} \overset{\text{D}}{=} \varepsilon_{tr} - \overline{\varepsilon}_{t\cdot}, \quad \overline{y}_{t\cdot} - \overline{y}_{\cdot\cdot} \overset{\text{D}}{=} \overline{\varepsilon}_{t\cdot} - \overline{\varepsilon}_{\cdot\cdot} + b_t - \overline{b}_{\cdot\cdot}$$

One can easily check that

$$\text{cov}(y_{tr} - \overline{y}_{t\cdot}, \overline{y}_{t\cdot} - \overline{y}_{\cdot\cdot}) = \text{cov}(\varepsilon_{tr} - \overline{\varepsilon}_{t\cdot}, b_t + \overline{\varepsilon}_{t\cdot} - \overline{b}_{\cdot} - \overline{\varepsilon}_{\cdot\cdot}) = 0,$$

so $y_{tr} - \overline{y}_{t\cdot} \perp\!\!\!\perp \overline{y}_{t\cdot} - \overline{y}_{\cdot\cdot}$ because all these variables are normal. Hence $SS_w \perp\!\!\!\perp SS_b$.

☐ Now $y_{tr} - \overline{y}_{t\cdot}$ does not depend on the b_t, so as usual

$$SS_w = \sum_{t=1}^{T}\sum_{r=1}^{R}(y_{tr} - \overline{y}_{t\cdot})^2 \overset{\text{D}}{=} \sum_{t=1}^{T}\sum_{r=1}^{R}(\varepsilon_{tr} - \overline{\varepsilon}_{t\cdot})^2 \sim \sigma^2\chi^2_{T(R-1)}.$$

The $\overline{y}_{t\cdot}$ are independent and $\overline{y}_{t\cdot} = \mu + b_t + \overline{\varepsilon}_{t\cdot} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma_b^2 + \sigma^2/R)$, so

$$SS_b = R\sum_{t=1}^{T}(\overline{y}_{t\cdot} - \overline{y}_{\cdot\cdot})^2 \overset{\text{D}}{=} R\sum_{t=1}^{T}(b_t + \overline{\varepsilon}_{t\cdot} - \overline{b}_t - \overline{\varepsilon}_{\cdot\cdot})^2 \sim R(\sigma_b^2 + \sigma^2/R)\chi^2_{T-1}.$$

Note II to Example 18

- Tests and confidence intervals for the ratio σ_b^2/σ^2 can be based on the $F_{T-1,T(R-1)}$ distribution of

$$\frac{\sigma^2}{\sigma^2 + R\sigma_b^2} \times \frac{SS_b/(T-1)}{SS_w/\{T(R-1)\}}. \qquad (6)$$

- One aspect of interest may be statements of uncertainty for the population mean μ, which is estimated by the overall sample average, $\overline{y}_{\cdot\cdot} = \mu + \overline{b}_{\cdot} + \overline{\varepsilon}_{\cdot\cdot}$. This has variance $\sigma_b^2/T + \sigma^2/(TR) = (\sigma^2 + R\sigma_b^2)/(TR)$, which is estimated unbiasedly by $SS_b/\{(T-1)TR\}$, independent of $\overline{y}_{\cdot\cdot}$, and confidence intervals are based on the t_{T-1} distribution of $(\overline{y}_{\cdot\cdot} - \mu)/[SS_b/\{(T-1)TR\}]^{1/2}$.
- Homogeneous variance across all blocks and normality can be checked using probability plots.

Nested model ANOVA

- Similar calculations for the nested model

$$y_{pvm} = \mu + b_p + e_{pv} + \varepsilon_{pvm}, \quad p = 1, \ldots, P, v = 1, \ldots, V, m = 1, \ldots, M,$$

give the ANOVA table below, in which each sum of squares is summed over p, v and m. Mean squares are formed by dividing sums of squares by their degrees of freedom.

- Below δ_b^2 and δ_e^2 are non-centrality parameters measuring differences among the b_p and e_{pv} when they are treated as fixed:

$$(P-1)\delta_b^2 = \sum_p (b_p - \overline{b}_{\cdot})^2, \quad P(V-1)\delta_e^2 = M \sum_{p,v} (e_{p,v} - \overline{e}_{p,\cdot})^2,$$

and when $b_p \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2) \perp\!\!\!\perp e_{pv} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$, we have $\mathrm{E}(\delta_b^2) = \sigma_b^2$, $\mathrm{E}(\delta_e^2) = \sigma_e^2$.

Term	df	Sum of squares	E(Mean square) when terms below random		
			ε	ε, e	ε, e, b
Between patients	$P-1$	$\sum(\overline{y}_{p\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2$	$VM\delta_b^2 + M\delta_e^2 + \sigma^2$	$VM\delta_b^2 + M\sigma_e^2 + \sigma^2$	$VM\sigma_b^2 + M\sigma_e^2 + \sigma^2$
Between visits within patients	$P(V-1)$	$\sum(\overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot})^2$	$M\delta_e^2 + \sigma^2$	$M\sigma_e^2 + \sigma^2$	$M\sigma_e^2 + \sigma^2$
Between measures within visits	$PV(M-1)$	$\sum(y_{pvm} - \overline{y}_{pv\cdot})^2$	σ^2	σ^2	σ^2

Analysis of variance

☐ Nested analysis of the blood pressure data:

```
summary( aov(dbp ~ patient/visno, data=blood.dia) )
               Df Sum Sq Mean Sq F value Pr(>F)
patient        24  23059   960.8  124.29 <2e-16 ***
patient:visno 375  39082   104.2   13.48 <2e-16 ***
Residuals     400   3092     7.7
```

☐ Likewise, crossed analysis of the chimpanzee data:

```
 summary( aov(log(y)~chimp+word,data=chimps) )
            Df Sum Sq Mean Sq F value  Pr(>F)
chimp        3   5.33   1.778   2.719  0.0642 .
word         9  45.69   5.077   7.765 1.5e-05 ***
Residuals   27  17.65   0.654
```

There are $C-1$ degrees of freedom for chimps, $W-1$ for words, and $(C-1)(W-1)$ for the residual.

☐ In both cases, we can use the ANOVA table to estimate the variance components and then perform **synthesis of variance**: e.g., how large would W need to be to distinguish the learning abilities of two chimps with probability 0.95?

Example: Blood pressure

☐ Solving the equations

$$\sigma^2 = 7.7, \quad M\sigma_e^2 + \sigma^2 = 104.2, \quad VM\sigma_b^2 + M\sigma_e^2 + \sigma^2 = 960.8,$$

gives (in units of millimeters of mercury, mmHg)

$$\widehat{\sigma} = 2.8, \quad \widehat{\sigma}_e = 6.9, \quad \widehat{\sigma}_b = 5.2,$$

so the largest variation is between different visits within patients, while that between measurements on a single visit is smallest.

Summary

☐ Components of variance ANOVA is easily performed for relatively simple linear, balanced data.

☐ Standard ANOVA tables have different interpretations, depending on which components of variance are taken to be random or fixed.

☐ Extensions are needed to deal with more complex settings, with unbalanced data, or with non-linear or non-normal errors.

☐ This leads to us to **mixed models**, i.e., models with both random and fixed parts.

Mixed models

- The term **mixed models** encompasses many different situations/models:
  - components of variance,
  - classical experimental design (split-plot designs, ...),
  - repeated measures,
  - longitudinal models,
  - multi-level models,
  - hierarchical models.
- Can subsume linear versions into the **linear mixed model**, which can be extended to nonlinear models, GLMs, ...

Linear mixed model

- The **linear mixed model** may be written as

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + Z_{n\times q}b_{q\times 1} + \varepsilon_{n\times 1}, \quad b \sim N_q(0, \Omega_b), \quad \varepsilon \sim \mathcal{N}_n(0, \Omega),$$

  where
  - β represents the **fixed effects**,
  - b represents the **random effects**, and
  - usually $\Omega = \sigma^2 I_n$.
- Equivalently,

$$y \mid b \ \sim \ \mathcal{N}_n(X\beta + Zb, \Omega), \quad b \ \sim \ \mathcal{N}_q(0, \Omega_b),$$

  which gives marginal response distribution

$$y \sim \mathcal{N}_n(X\beta, Z\Omega_b Z^{\mathrm{T}} + \Omega), \quad Z\Omega_b Z^{\mathrm{T}} + \Omega = \sigma^2\Delta^{-1}(\psi),$$

  say, with ψ denoting the vector of distinct variance ratios appearing in Δ^{-1} (e.g., σ_b^2/σ^2 in Example 18).
- Although Ω is often diagonal, $Z\Omega_b Z^{\mathrm{T}}$ is not, so inverting $Z\Omega_b Z^{\mathrm{T}} + \Omega$ involves $O(n^3)$ flops in general, and we should try and avoid working with Δ.

Naive maximum likelihood estimation

- The log likelihood based on $f(y;\beta,\sigma^2,\psi)$ is

$$\ell(\beta,\sigma^2,\psi) \equiv -\frac{1}{2\sigma^2}(y-X\beta)^{\mathrm{T}}\Delta(y-X\beta) - \frac{n}{2}\log\sigma^2 + \tfrac{1}{2}\log|\Delta|,$$

where $\Delta \equiv \Delta(\psi)$. For known ψ (and hence known Δ) the MLEs of β and σ^2 are

$$\widehat{\beta}_\psi = (X^{\mathrm{T}}\Delta X)^{-1}X^{\mathrm{T}}\Delta y, \quad \widehat{\sigma}^2_\psi = n^{-1}(y-X\widehat{\beta}_\psi)^{\mathrm{T}}\Delta(y-X\widehat{\beta}_\psi),$$

so the profile log likelihood for ψ is

$$\ell_{\mathrm{p}}(\psi) \equiv -\tfrac{1}{2}n\log\widehat{\sigma}^2_\psi + \tfrac{1}{2}\log|\Delta(\psi)|.$$

- We could maximize $\ell_{\mathrm{p}}(\psi)$ to estimate ψ, and then obtain maximum likelihood estimates $\widehat{\beta}_{\widehat{\psi}}$ and $\widehat{\sigma}^2_{\widehat{\psi}}$. This involves maximization of $\ell_{\mathrm{p}}(\psi)$, either by **Newton–Raphson** (quadratic convergence, may be unstable) or using the **EM algorithm** (linear convergence but more stable) (later).
- We also want inference on b ...

Improved approach

- Let $\tilde{b}$ denote the MLE of b for fixed β (and ψ). Then

$$\begin{aligned} f(y;\beta,\sigma^2,\psi) &= \int f(y\mid b;\beta,\sigma^2,\psi)f(b;\sigma^2,\psi)\,\mathrm{d}b \\ &= f(y,\tilde{b};\beta,\sigma^2,\psi)\times\frac{(2\pi)^{p/2}}{|Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|^{1/2}} \\ &\propto \frac{f(y\mid\tilde{b};\beta,\sigma^2,\psi)f(\tilde{b}\mid\sigma^2,\psi)}{|Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|^{1/2}}, \end{aligned}$$

so (apart from additive constants) $-2\log f(y;\beta,\sigma^2,\psi)$ equals

$$(y-X\beta-Z\tilde{b})^{\mathrm{T}}\Omega^{-1}(y-X\beta-Z\tilde{b}) + \tilde{b}^{\mathrm{T}}\Omega_b^{-1}\tilde{b} + \log\{|\Omega||\Omega_b||Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}|\}.$$

- The first two (quadratic) terms here depend on β and b, so given ψ and σ^2 we can find $\widehat{\beta}_\psi$ and $\tilde{b}(\widehat{\beta},\psi)$ explicitly, and thus obtain $\ell_{\mathrm{p}}(\psi)$.
- By noting that

$$f(b\mid y;\beta,\sigma^2,\psi) = f(y\mid b;\beta,\sigma^2,\psi)f(b;\sigma^2,\psi)/f(y;\beta,\sigma^2,\psi)$$

and taking logs, we obtain

$$b\mid y\sim\mathcal{N}_q\left\{\tilde{b},(Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1})^{-1}\right\}, \quad \tilde{b} = \left(Z^{\mathrm{T}}\Omega^{-1}Z+\Omega_b^{-1}\right)^{-1}Z^{\mathrm{T}}\Omega^{-1}(y-X\beta).$$

Note on improved approach

- Suppressing the parameters β, σ^2 and ψ for now, we write the log integrand in

$$f(y) = \int f(y, b)\,\mathrm{d}b = \int f(y \mid b) f(b)\,\mathrm{d}b$$

in the form

$$\log f(y, b) = \log f(y, \tilde{b}) - \tfrac{1}{2}(b - \tilde{b})^{\mathrm{T}} H(\tilde{b})(b - \tilde{b}),$$

where the linear term of the Taylor series equals zero, because it is evaluated at the maximising value $\tilde{b}$, and the given Taylor series is exact because the log likelihood is quadratic.

- On ignoring terms not involving b we have

$$-2\log f(y, b) = -2\log f(y \mid b) - 2\log f(b) \equiv (y - X\beta - Zb)^{\mathrm{T}}\Omega^{-1}(y - X\beta - Zb) + b^{\mathrm{T}}\Omega_b^{-1}b,$$

so

$$H(b) \equiv H = Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}$$

does not depend on b, and thus

$$\begin{aligned} f(y) &= f(y, \tilde{b}) \int \exp\left\{-\tfrac{1}{2}(b - \tilde{b})^{\mathrm{T}} H (b - \tilde{b})\right\} \mathrm{d}b \\ &= f(y, \tilde{b}) \times (2\pi)^{q/2} |H|^{-1/2} \\ &= f(y, \tilde{b}) \times \frac{(2\pi)^{q/2}}{|Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}|^{1/2}}, \end{aligned}$$

as announced; the integral equals the normalising constant for a Gaussian density with variance matrix H^{-1}.

Reminder: Matrix inversion formulae

- Assume that any inverses needed below exist.
- The (Sherman–Morrison)–Woodbury formula

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

is easily checked by multiplying the formula above by $A + BCD$.

- For the partitioned matrix A and its correspondingly partitioned inverse,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix},$$

we have

$$A^{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}, \qquad A^{12} = -A_{11}^{-1}A_{12}A^{22},$$
$$A^{21} = -A_{22}^{-1}A_{21}A^{11}, \qquad A^{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1},$$

as is also easily checked.

Inference on β

- Since
$$y \sim \mathcal{N}_n(X\beta, Z\Omega_b Z^{\mathrm{T}} + \Omega),$$
weighted least squares gives
$$\widehat{\beta} = \{X^{\mathrm{T}}(Z\Omega_b Z^{\mathrm{T}} + \Omega)^{-1}X\}^{-1}X^{\mathrm{T}}(Z\Omega_b Z^{\mathrm{T}} + \Omega)^{-1}y,$$
with
$$\widehat{\beta} \sim \mathcal{N}_p\left[\beta, \{X^{\mathrm{T}}(Z\Omega_b Z^{\mathrm{T}} + \Omega)^{-1}X\}^{-1}\right],$$
where in general we need $O(n^3)$ flops to invert the $n \times n$ matrix $Z\Omega_b Z^{\mathrm{T}} + \Omega$.
- For cheaper calculation of $\mathrm{var}(\widehat{\beta})$, we use the inversion formulae and obtain
$$\begin{pmatrix} \mathrm{var}(\widehat{\beta})_{p\times p} & \cdot \\ \cdot & \cdot \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Omega^{-1}X & X^{\mathrm{T}}\Omega^{-1}Z \\ Z^{\mathrm{T}}\Omega^{-1}X & Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1} \end{pmatrix}^{-1}_{(p+q)\times(p+q)},$$
which involves only $O\{n(p+q)^2\}$ flops, as Ω is usually diagonal.
- Note that $\mathrm{var}(b \mid y) = (Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1})^{-1}$ can be obtained as a by-product.
- In practice these formulae would be evaluated at the MLEs $\widehat{\sigma}^2$ and $\widehat{\psi}$ and used to compute confidence intervals etc. for elements of β.

Inference on random effects

- **Conventional terminology**: we **estimate** parameters β and **predict** random variables b.
- To find the best predictor $\tilde{b}(y)$ of b we minimise
$$\mathrm{E}_{b,y}\left[\left\{\tilde{b}(y) - b\right\}^{\mathrm{T}}\left\{\tilde{b}(y) - b\right\}\right],$$
which gives $\tilde{b}(y) = \mathrm{E}(b \mid y)$, with (Woodbury formula):
$$\begin{aligned} \mathrm{E}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}(y - X\beta), \\ \mathrm{var}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1}. \end{aligned}$$
- Replace parameters β, σ^2, ψ by estimates to get **best linear unbiased predictor (BLUP)** $\tilde{b}$ and its estimated variance.
- Residuals
$$\begin{aligned} y - X\widehat{\beta} &= Z\tilde{b} + y - X\widehat{\beta} - Z\tilde{b} \\ &= Z\tilde{b} + \left\{I_n - Z\left(Z^{\mathrm{T}}\widehat{\Omega}^{-1}Z + \widehat{\Omega}_b^{-1}\right)^{-1} Z^{\mathrm{T}}\widehat{\Omega}^{-1}\right\}\left(y - X\widehat{\beta}\right), \end{aligned}$$
split into two parts, with $Z\tilde{b}$ attributable to random effects, and the second the usual residual $y - X\widehat{\beta}$ shrunk towards zero; this estimates ε.

Note on conditional mean and variance

- First we write
$$\tilde b(y) - b = \tilde b(y) - \mathrm{E}(b \mid y) + \mathrm{E}(b \mid y) - b,$$
expand $\{\tilde b(y) - b\}^{\mathrm T}\{\tilde b(y) - b\}$ and take expectation over b conditional on y to get
$$\mathrm{E}\left[\{\tilde b(y) - b\}^{\mathrm T}\{\tilde b(y) - b\} \mid y\right] = \left\{\tilde b(y) - \mathrm{E}(b \mid y)\right\}^{\mathrm T}\left\{\tilde b(y) - \mathrm{E}(b \mid y)\right\} + \mathrm{var}(b \mid y),$$
which is minimised when $\tilde b(y) = \mathrm{E}(b \mid y)$. Any other choice will give a larger expectation when we take E_y, so this is optimal.
- To obtain $\mathrm{E}(b \mid y)$, we note that
$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathcal{N}_{n+q}\left\{\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega + Z\Omega_b Z^{\mathrm T} & Z\Omega_b \\ \Omega_b Z^{\mathrm T} & \Omega_b \end{pmatrix}\right\},$$
so using standard formulae for conditional normal distributions, we have
$$\begin{aligned} \mathrm{E}(b \mid y) &= \Omega_b Z^{\mathrm T}\left(\Omega + Z\Omega_b Z^{\mathrm T}\right)^{-1}(y - X\beta), \\ \mathrm{var}(b \mid y) &= \Omega_b - \Omega_b Z^{\mathrm T}\left(\Omega + Z\Omega_b Z^{\mathrm T}\right)^{-1} Z\Omega_b. \end{aligned}$$
- The Woodbury formula applied to the conditional variance gives
$$\mathrm{var}(b \mid y) = \left(Z^{\mathrm T}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1}$$
as required.
- For the conditional mean we apply the Woodbury formula to $(\Omega + Z\Omega_b Z^{\mathrm T})^{-1}$ and get
$$\begin{aligned} \mathrm{E}(b \mid y) &= \Omega_b Z^{\mathrm T}\left\{\Omega^{-1} - \Omega^{-1}Z\left(\Omega_b^{-1} + Z^{\mathrm T}\Omega^{-1}Z\right)^{-1} Z^{\mathrm T}\Omega^{-1}\right\}(y - X\beta) \\ &= \Omega_b\left\{I_q - Z^{\mathrm T}\Omega^{-1}Z\left(\Omega_b^{-1} + Z^{\mathrm T}\Omega^{-1}Z\right)^{-1}\right\} Z^{\mathrm T}\Omega^{-1}(y - X\beta) \\ &= \Omega_b\left\{\Omega_b^{-1}\left(\Omega_b^{-1} + Z^{\mathrm T}\Omega^{-1}Z\right)^{-1}\right\} Z^{\mathrm T}\Omega^{-1}(y - X\beta), \end{aligned}$$
as required, where we wrote the term in braces in the second line as
$I - B(A + B)^{-1} = A(A + B)^{-1}$, with $A = \Omega_b^{-1}$ and $B = Z^{\mathrm T}\Omega^{-1}Z$.

Example

Example 19 (One-way layout) *Work out the details for the unbalanced one-way layout model*
$$y_{ij} = \mu + b_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, q,$$
in which
$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2) \quad \perp\!\!\!\perp \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Note to Example 19

☐ We have

$$\Omega = \sigma^2 I_n, \quad \Omega_b = \sigma_b^2 I_q, \quad X = 1_n, \quad Z = \begin{pmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{n_q} \end{pmatrix},$$

where $n = n_1 + \cdots + n_q$.

☐ We first need to compute $\widehat{\beta} = (X^{\mathrm{T}}\Upsilon X)^{-1}X^{\mathrm{T}}\Upsilon y$, where $\sigma^2\Upsilon^{-1} = Z\Omega_b Z^{\mathrm{T}} + \Omega$. Letting $I_i = I_{n_i}$, $1_i = 1_{n_i}$ and $J_i = 1_i 1_i^{\mathrm{T}}$ for shorthand, we see that $Z\Omega_b Z^{\mathrm{T}} + \Omega$ is block diagonal with blocks

$$\sigma^2 I_i + \sigma_b^2 J_i = \sigma^2(I_i + \psi J_i), \quad i = 1, \ldots, q, \quad \psi = \sigma_b^2/\sigma^2,$$

and hence that Υ is block diagonal with blocks

$$I_i - \frac{\psi}{1 + n_i\psi} J_i,$$

so

$$X^{\mathrm{T}}\Upsilon X = \sum_{i=1}^{q} 1_i^{\mathrm{T}}\left(I_i - \frac{\psi}{1+n_i\psi}J_i\right)1_i = \sum_{i=1}^{q}\left(n_i - \frac{n_i^2\psi}{1+n_i\psi}\right) = \sum_{i=1}^{q}\frac{n_i}{1+n_i\psi}.$$

Similarly we find that

$$X^{\mathrm{T}}\Upsilon y = \sum_{i=1}^{q} 1_i\left(I_i - \frac{\psi}{1+n_i\psi}J_i\right)y_i = \sum_{i=1}^{q}\frac{n_i\overline{y}_{i\cdot}}{1+n_i\psi}, \quad \widehat{\mu}_\psi = \frac{\sum_{i=1}^{q} n_i\overline{y}_{i\cdot}/(1+n_i\psi)}{\sum_{i=1}^{q} n_i/(1+n_i\psi)}.$$

☐ Note that $\widehat{\mu}_\psi = \overline{y}_{\cdot\cdot}$ if all the n_i are equal. If $\psi = 0$, then $\widehat{\mu}_\psi = \sum_{i,j} y_{ij}/\sum_i n_i$ equals the grand mean, whereas if $\psi \to \infty$, $\widehat{\mu}_\psi \to q^{-1}\sum_i \overline{y}_{i\cdot}$. In the first case, there is no variation among the b_i, and so the data should be treated as a simple random sample of size $\sum_i n_i$, whereas in the second case, the variation of the y_{ij} around b_i tends to zero, so the replicates cannot be used and the best estimate of μ is the average of the group means.

☐ It is easy to check using calculations similar to the above that

$$(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1})^{-1} = \sigma^2\mathrm{diag}\{\psi/(1+n_i\psi)\}, \quad [Z^{\mathrm{T}}\Omega^{-1}(y - X\widehat{\beta})]_i = n_i(\overline{y}_{i\cdot} - \widehat{\mu}_{\widehat{\psi}})/\sigma^2,$$

and hence that the ith element of $\tilde{b}$ and its estimated variance can be written as

$$\tilde{b}_i = \frac{\overline{y}_{i\cdot} - \widehat{\mu}_{\widehat{\psi}}}{1 + \widehat{\sigma}^2/(n_i\widehat{\sigma}_b^2)}, \quad \frac{1}{1/\widehat{\sigma}_b^2 + n_i/\widehat{\sigma}^2},$$

so $\overline{y}_{i\cdot} - \widehat{\mu}_{\widehat{\psi}}$ is shrunk towards zero by an amount that depends on the estimated variance ratio. The shrinkage will be considerable if $\widehat{\sigma}^2/n_i \gg \widehat{\sigma}_b^2$, corresponding to large variation in the group averages owing to individual variances compared to the variation between groups. The data are then almost a simple random sample of size n, so strong shrinkage is not surprising.

☐ Note also that $\mathrm{var}(\tilde{b}_i \mid y) \to 0$ when $\sigma_b^2 \to 0$, $\sigma^2 \to 0$, or $n_i \to \infty$. In the first case, there is no variation between groups, and hence $b_i = 0$ with probability one. In the second two cases, the value of b_i is known exactly, because variation around it is negligible. Thus consistent inference for b_i is impossible when σ_b^2 and σ^2 take positive values: even if $q \to \infty$, the amount of information on any given b_i does not accumulate unless $n_i \to \infty$, and this is rarely the case.

REML

- Problems with the usual MLEs:
  - the MLEs $\widehat{\sigma}^2$ and $\widehat{\psi}$ are downwardly biased (no adjustment for estimation of β);
  - variance ratios $\psi_r \geq 0$, so maybe $\widehat{\psi}_r = 0$, making usual asymptotic properties fail;
  - n can be very large in applications, raising computational issues.
- **Restricted maximum likelihood (REML) estimation** is often used to reduce the bias:
  - a form of **marginal likelihood**, i.e., based on the marginal density of a statistic $s = s(y)$;
  - but can also be derived as a **conditional likelihood**, based on the conditional density of y given a statistic $s(y)$.
- REML maximised using Newton–Raphson or EM algorithms.

Lemma 20 *A restricted log likelihood for σ^2 and ψ is*

$$\ell(\widehat{\beta}_\psi, \sigma^2, \psi) + \frac{p}{2}\log\sigma^2 - \tfrac{1}{2}\log|X^{\mathrm{T}}\Delta X|$$

i.e.,

$$\tfrac{1}{2}\log|\Delta| - \tfrac{1}{2}\log|X^{\mathrm{T}}\Delta X| - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Delta(y - X\widehat{\beta}_\psi) - \frac{n-p}{2}\log\sigma^2.$$

Note to Lemma 20

- Suppose $f(y;\alpha,\beta)$ depends on two parameters, that interest is focused on α, and that for fixed α there is a sufficient statistic s_α for β. Then

$$f(y;\alpha,\beta) = f(y \mid s_\alpha;\alpha) f(s_\alpha;\alpha,\beta),$$

  and since the first density on the right is a proper conditional density not depending on β, we can base inference for α on

$$\log f(y \mid s_\alpha;\alpha) = \log f(y;\alpha,\beta) - \log f(s_\alpha;\alpha,\beta).$$

- In the normal mixed model we take $\alpha = (\sigma^2,\psi)$. If all the variance parameters are fixed, then $s_\alpha = \widehat{\beta}_\alpha = (X^{\mathrm{T}}\Delta X)^{-1}X^{\mathrm{T}}\Delta y$ is sufficient for β; its distribution is $\mathcal{N}_p\{\beta, \sigma^2(X^{\mathrm{T}}\Delta X)^{-1}\}$.
- Apart from constants, the logarithm of the required conditional density is therefore

$$\begin{aligned}\log f(y;\sigma^2,\psi,\beta) - \log f(\widehat{\beta}_\psi;\sigma^2,\psi,\beta) \quad &\equiv \quad -\frac{n}{2}\log\sigma^2 + \tfrac{1}{2}\log|\Delta| - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}\Delta(y - X\beta)\\ &\quad + \frac{p}{2}\log\sigma^2 - \tfrac{1}{2}\log|X^{\mathrm{T}}\Delta X|\\ &\quad + \frac{1}{2\sigma^2}(\widehat{\beta}_\psi - \beta)^{\mathrm{T}}X^{\mathrm{T}}\Delta X(\widehat{\beta}_\psi - \beta),\end{aligned}$$

  which reduces to the given form on writing

$$y - X\beta = (y - X\widehat{\beta}_\psi) + X(\widehat{\beta}_\psi - \beta)$$

  in the first quadratic term and expanding out, noting that the cross-product vanishes because

$$\begin{aligned}(\widehat{\beta}_\psi - \beta)^{\mathrm{T}}X^{\mathrm{T}}\Delta(y - X\widehat{\beta}_\psi) &= (\widehat{\beta}_\psi - \beta)^{\mathrm{T}}X^{\mathrm{T}}\Delta\{y - X(X^{\mathrm{T}}\Delta X)^{-1}X^{\mathrm{T}}\Delta y\}\\ &= (\widehat{\beta}_\psi - \beta)^{\mathrm{T}}\{X^{\mathrm{T}}\Delta y - X^{\mathrm{T}}\Delta X(X^{\mathrm{T}}\Delta X)^{-1}X^{\mathrm{T}}\Delta y\} = 0.\end{aligned}$$

Example: Rat growth

Example 21 (Rat growth data)

- ☐ *Write*

$$y_{jt} = \beta_0 + b_{j0} + (\beta_1 + b_{j1})x_{jt} + \varepsilon_{jt}, \quad t = 1, \dots 5, j = 1, \dots, 30,$$

where the random variables (b_{j0}, b_{j1}) *have a joint normal distribution with mean vector zero and unknown variance matrix and the* $\varepsilon_{jt} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. *In matrix terms,*

$$\begin{pmatrix} y_{j1} \\ \vdots \\ y_{j5} \end{pmatrix} = \begin{pmatrix} 1 & x_{j1} \\ \vdots & \vdots \\ 1 & x_{j5} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & x_{j1} \\ \vdots & \vdots \\ 1 & x_{j5} \end{pmatrix} \begin{pmatrix} b_{j0} \\ b_{j1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{j1} \\ \vdots \\ \varepsilon_{j5} \end{pmatrix}, \quad j = 1, \dots, 30;$$

the overall model with $n = 150$ *is obtained by stacking these expressions.*

- ☐ *We set* $(x_{j1}, \dots, x_{j5}) = (0, \dots, 4)$, *so that* β_0 *is the mean weight in week 1.*
- ☐ $p = 2$ *parameters;* $q = 60$ *since two random variables per rat.*

Example: Rat growth

```
> rat.growth
   rat week   y
1    1    0 151
2    1    1 199
3    1    2 246
4    1    3 283
5    1    4 320
6    2    0 145
...
> fit.reml <- lme(fixed= y~week, random=~week|rat, data=rat.growth)
> summary(fit.reml)
Linear mixed-effects model fit by REML
 Data: rat.growth
      AIC      BIC    logLik
  1096.58 1114.563 -542.2899

Random effects:
 Formula: ~week | rat
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 10.932986 (Intr)
week         3.534747 0.184
Residual     5.817426

Fixed effects: y ~ week
                Value Std.Error  DF  t-value p-value
(Intercept) 156.05333 2.1589786 119 72.28109       0
week         43.26667 0.7275228 119 59.47122       0
 Correlation:
     (Intr)
week 0.007
```

Example: Rat growth

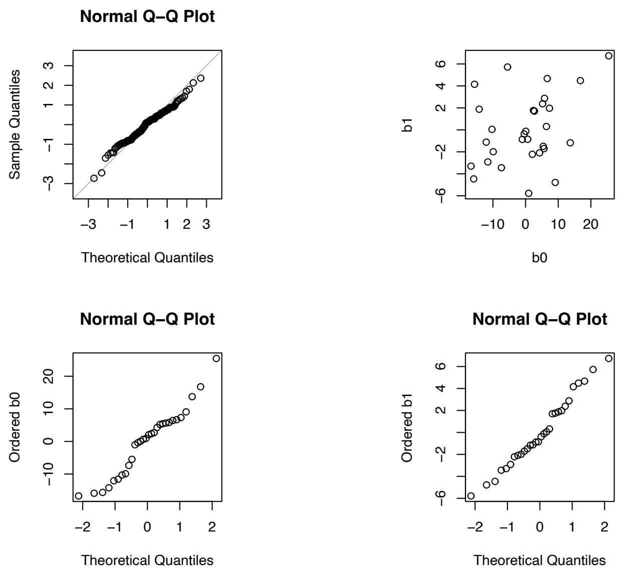
Results from fit of mixed model to rat growth data, using REML. Values in parentheses are for ML fit. In each case $\widehat{\sigma}^2 = 5.82^2$.

Parameter	Fixed		Random	
	Estimate	Standard error	Variance	Correlation
Intercept	156.05	2.16 (2.13)	10.93^2 (10.71^2)	
Slope	43.27	0.73 (0.72)	3.53^2 (3.46^2)	0.18 (0.19)

- ☐ REML estimates of Ω_b slightly larger than ML estimates, but effect is small since $p = 2$.
- ☐ Estimated mean weight in week 1 is 156, but SD of individual rats around this is 11.
- ☐ Correlation between slope and intercept is small but positive: initially heavier rats tend to gain weight faster.
- ☐ Variation around individual slopes is given by $\widehat{\sigma}$, smaller than for the intercept variance.
- ☐ Shrinkage of intercept estimates, shown on next page, is small in this case.
- ☐ Residuals look acceptably normal.

Example: Rat growth

Residuals and random effects



Comments

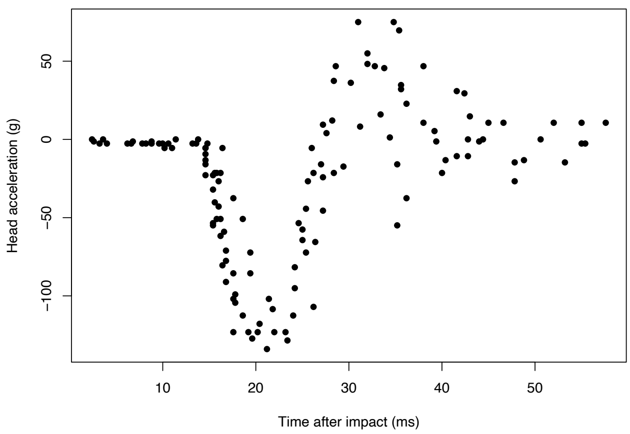
- ☐ Testing for non-zero variance components (e.g., $\psi = 0$ in Example 19) involves tests on the boundary of the parameter space, which have nasty asymptotic properties: if $\psi = 0$, then in that example, the likelihood ratio statistic for testing $\psi = 0$ satisfies $W \dot\sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ as $n \to \infty$, meaning that
$$\mathrm{P}_0(W = 0) = \tfrac{1}{2}, \quad \mathrm{P}_0(W > w) = \tfrac{1}{2}\mathrm{P}(\chi_1^2 > w), \quad w > 0.$$
  Unfortunately,
  - $\mathrm{P}_0(W = 0)$ can be very different from $\frac{1}{2}$ even in large samples, and
  - in more complex problems, the limiting distribution can be much more complex.
- ☐ Sometimes clearer to write a mixed model in **multi-level model** form
$$y = X\beta + Z_L b_L + \cdots + Z_0 b_0,$$
  where the $q_l \times 1$ vectors b_l are all mutually independent with means zero and variance matrices Ω_l, so $Y \sim \mathcal{N}_n(X\beta, \sum_{l=0}^{L} Z_l \Omega_l Z_l^{\mathrm{T}})$, where $Z_0 = I_n$, $b_0 = \varepsilon$ and $\Omega_0 = \sigma^2 I_n$.
- ☐ The same basic approaches apply in **nonlinear mixed models** and **generalized linear mixed models (GLMMs)**, but integrals appear everywhere and have to be approximated numerically, leading to heavier computational burdens (later).

Motivation

- Normal linear model has two main aspects:
  - **systematic variation**, $\mathrm{E}(y) = \mu$, and $\mu = X\beta$ with parameters β;
  - **stochastic variation**, $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$.
- Can relax the stochastic assumption using other distributions or second-order assumptions, but still have parametric model for the systematic part.
- Often want to relax systematic part for more flexible models, for
  - exploratory data analysis — 'will a linear model be adequate?'
  - confirmatory data analysis — 'I've fitted a linear model, is it adequate?'
  - general modelling — 'the data are too complex to expect a simple parametric model to work, so what can I do?'
  - semiparametric modelling — 'I will use a parametric model for the effects of interest, but can I model nuisance effects more flexibly?'
- Introduce **non/semi-parametric** regression models—widely used in applications.

Example: Motorcycle data

Measurements of head acceleration (g) at time after impact (ms) in a simulated motorcycle accident, used to test crash helmets:

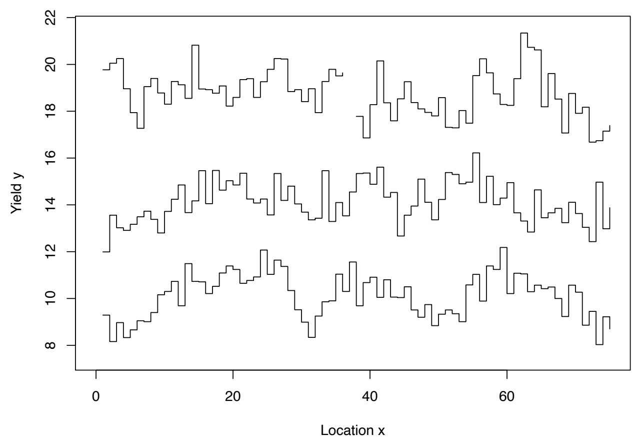


Example: Spring barley data

Plot yield at harvest for 75 varieties of spring barley sown in 3 blocks each of 75 plots:

Location x	Block 1		Block 2		Block 3	
	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example: Spring barley data



Yield as a function of location for the three blocks, with yields for blocks 2 and 3 offset by the addition of 4 and of 7 respectively. Value 37 in block 3 is missing.

Comments

- The motorcycle data are pairs $(x_1, y_1), \ldots, (x_n, y_n)$, with

$$y_j \stackrel{\text{ind}}{\sim} (\mu(x_j), \sigma^2), \quad j = 1, \ldots, n,$$

  with μ unspecified and $x_1, \ldots, x_n \in \mathcal{X}$, say (ignoring the heteroscedasticity). We might aim to estimate the function $\mu(\cdot)$ using a **scatterplot smoother**.
- For the spring barley data we could envisage an **additive model**

$$y_{bx} \stackrel{\text{ind}}{\sim} (\beta_{v(b,x)} + \mu_b(x), \sigma^2), \quad b = 1, 2, 3, x = 1, \ldots, 75,$$

  where
  - $\beta_{v(b,x)}$ denotes mean yield for the variety $v(b, x)$ planted at location x of block b, and
  - $\mu_b(x)$ represents smooth variation in the mean yield in block b, as fertility changes with location x.

  In this case we are mostly interested in the variety effects $\beta_1, \ldots, \beta_{75}$, but need to account for changes in the (nuisance functions) μ_b.
- More generally we might have random effects (the β_v?) and/or non-linear models and/or non-Gaussian responses ...

Polynomial regression

- Fit polynomial of degree $p - 1$, i.e.,

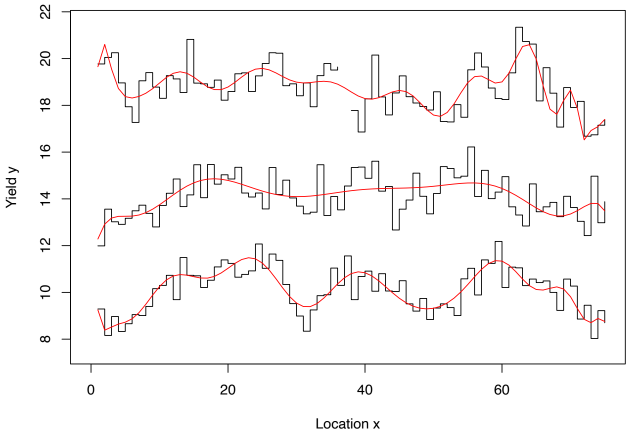
$$\mu(x_j) = \beta_0 + \beta_1 x_j + \cdots + \beta_{p-1} x_j^{p-1},$$

  and choose $\beta_0, \ldots, \beta_{p-1}$ to minimise the sum of squares

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 = \sum_{j=1}^{n} \left\{ y_j - (\beta_0 + \beta_1 x_j + \cdots + \beta_{p-1} x_j^{p-1}) \right\}^2,$$

  giving $\widehat{\beta}_{p\times 1} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y$, where (j, i) element of $n \times p$ matrix X is x_j^{i-1}.
- Comments:
  - easily copes with missing values/unequally spaced observations;
  - use orthogonal polynomials to avoid numerical problems if n, k large;
  - sensitivity to observations at extremities of series often leads to poor fit;
  - usually doesn't work well because polynomials are too restrictive.

Spring barley data and polynomial fits



Yield as a function of location for the three blocks, with yields for blocks 2 and 3 offset by the addition of 4 and of 7 respectively, with fitted polynomials of degrees 20, 10 and 50.

Local polynomial regression

- ☐ Idea is estimate $\mu(x)$ near $x = x_0$ by fitting a low-order polynomial to the data nearest to x_0.
- ☐ Use kernel function $w(\cdot)$ with bandwidth h to give weights

$$w_h(x - x_0) = h^{-1} w\{(x - x_0)/h\}$$

that downweight observations far from x_0, and minimise weighted sum of squares

$$\sum_{j=1}^{n} w_h(x_j - x_0) \left[y_j - \left\{\beta_0 + \beta_1(x_j - x_0) + \cdots + \beta_{p-1}(x_j - x_0)^{p-1}\right\}\right]^2,$$

giving

$$\widehat{\beta}(x_0) = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W y, \quad \widehat{\mu}(x_0) = \widehat{\beta}_0(x_0),$$

where y and X are as before and the $n \times n$ diagonal matrix W contains the weights.

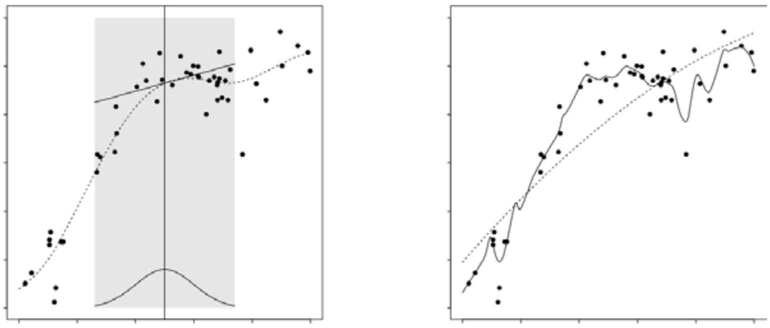
- ☐ Refit for numerous x_0, and interpolate the fitted values to estimate $\mu(x)$.
- ☐ Since

$$\widehat{\mu}(x_0) = \widehat{\beta}_0(x_0) = (1, 0, \ldots, 0)^{\mathrm{T}} \widehat{\beta}(x_0)$$

is a linear function of y, the vector of fitted values can be written as $\widehat{\mu}_{n\times 1} = S_h y$ after setting x_0 successively equal to $x_1, \ldots, x_n$.

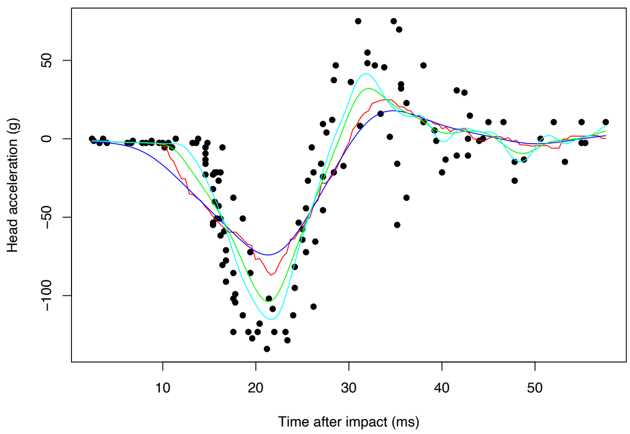
Local linear smoother

Left: 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 $\mu(x)$.
Right: local linear smoothers with $h = 0.2$ (solid) and $h = 5$ (dots).



Motorcycle data

Data with four local constant (Nadaraya–Watson, $p = 1$) estimates: uniform kernel, $h = 10$ (red); normal kernel, $h = 10$ (blue); normal kernel, $h = 5$ (green); normal kernel, $h = 3$ (cyan).



Comments

- ☐ Local polynomial fitting provides linear smoothers that depend on:
  - a kernel function w (not really important, provided edges smooth);
  - a bandwidth h (important, chosen by cross-validation, AIC or similar);
  - degree $p-1$ of polynomial, with $p=2$ (local linear smoother) most common.
- ☐ Can robustify by downweighting observations with large residuals in initial fit.
- ☐ **Lowess** (**lo**cally **we**ighted **s**catterplot **s**moother) uses nearest neighbourhood smoother, which smooths the 2/3 of the x_j closest to x_0—equivalent to a varying bandwidth.
- ☐ Extends to other models using **local likelihood estimation**, where we maximise

$$\ell(\theta; x_0) = \sum_{j=1}^{n} w_h(x_j - x_0)\, \ell_j(\beta; x_0)$$

to get $\widehat{\beta}(x_0)$ using iterative weighted least squares (later) and a suitable definition of $\mathrm{AIC_c}(h)$.

Function Estimation slide 148

Loss, risk and admissibility

- ☐ A **loss function** $l(\widehat{\mu}, \mu)$ represents cost of estimating μ by $\widehat{\mu}$.
- ☐ The average cost of estimating μ by $\widehat{\mu}$ is measured by the **risk function**

$$R_{\widehat{\mu}}(\mu) = \mathrm{E}\left\{l(\widehat{\mu}, \mu)\right\}.$$

- ☐ An estimator $\widehat{\mu}$ of μ is **inadmissible** if another estimator $\tilde{\mu}$ exists such that

$$R_{\widehat{\mu}}(\mu) \geq R_{\tilde{\mu}}(\mu),$$

with strict inequality for some μ. Then $\tilde{\mu}$ is never worse than $\widehat{\mu}$, and sometimes better than it, so $\widehat{\mu}$ should not be used.

- ☐ An estimator is **admissible** if it is not inadmissible.
- ☐ There are many possibilities, but for simplicity we consider **squared error loss**

$$l(\widehat{\mu}, \mu) = (\widehat{\mu} - \mu)^{\mathrm{T}}(\widehat{\mu} - \mu),$$

whose corresponding risk is the **mean square error**

$$R_{\widehat{\mu}}(\mu) = \mathrm{E}\left\{(\widehat{\mu} - \mu)^{\mathrm{T}}(\widehat{\mu} - \mu)\right\}.$$

Stein's theorem

Theorem 22 (Stein) *If $y \sim \mathcal{N}_n(\mu, I_n)$ and $n > 3$, then the maximum likelihood estimator $\widehat{\mu}$ of μ is inadmissible in terms of mean square error: the shrinkage estimator*

$$\tilde{\mu} = \overline{y}1_n + \left(1 - \frac{b}{w}\right)(y - \overline{y}1_n), \quad b \geq 0,$$

where $w = \sum_j (y_j - \overline{y})^2$ and $\overline{y} = n^{-1}\sum y_j$, has risk

$$R_{\tilde{\mu}}(\mu) = n + b\{b - 2(n-3)\}\,\mathrm{E}(w^{-1}),$$

and thus $R_{\widehat{\mu}}(\mu) > R_{\tilde{\mu}}(\mu)$ for any $\mu \in \mathbb{R}^n$.

- ☐ As $b \to 0$, $\tilde{\mu} \to y$, and as $b \to w$, $\tilde{\mu} \to \overline{y}1_n$.
- ☐ As $\mathrm{E}(w^{-1}) > 0$, $\widehat{\mu}$ is inadmissible if $n > 3$: it is always better to shrink y towards $\overline{y}1_n$.
- ☐ The minimum of $R_{\tilde{\mu}}(\mu)$ is $n - (n-3)^2\mathrm{E}(w^{-1})$ when $b = n - 3$.
- ☐ If $\mu_1 = \cdots = \mu_n$, then $\mathrm{E}(w^{-1}) = (n-3)^{-1}$ and $R_{\tilde{\mu}}(\mu) = 3$: the decrease in risk can be enormous if the μ_j are very similar, but it is smaller if they are very variable.
- ☐ This assumes it makes sense to cumulate risk over all μ_j; there may be no improvement for estimating individual parameters.

Proof of Theorem 22

☐ We first note that the MLE of μ is y, and that its risk is

$$R_{\widehat{\mu}}(\mu) = \mathrm{E}\{(\widehat{\mu}-\mu)^{\mathrm{T}}(\widehat{\mu}-\mu)\} = \mathrm{E}\{(y-\mu)^{\mathrm{T}}(y-\mu)\} = \mathrm{E}\left\{\sum_{j=1}^{n}(y_j-\mu_j)^2\right\} = \sum_{j=1}^{n}\mathrm{var}(y_j) = n.$$

☐ Now

$$\begin{aligned}(\tilde{\mu}-\mu)^{\mathrm{T}}(\tilde{\mu}-\mu) &= \sum(\tilde{\mu}_j-\mu_j)^2\\ &= \sum_{j=1}^{n}\{\overline{y}+(1-b/w)(y_j-\overline{y})-\mu_j\}^2\\ &= \sum_{j=1}^{n}\{y_j-\mu_j-b(y_j-\overline{y})/w\}^2\\ &= \sum_{j=1}^{n}(y_j-\mu_j)^2-2\frac{b}{w}\sum_{j=1}^{n}(y_j-\mu_j)(y_j-\overline{y})+\frac{b^2}{w^2}\sum_{j=1}^{n}(y_j-\overline{y})^2. \qquad (7)\end{aligned}$$

The first and third terms have expectations n and $b^2\mathrm{E}(w^{-1})$, so we must deal with the second.

☐ Consider $\mathrm{E}\{(y_1-\mu_1)h_1(y_1)\}$, where h_1 is a sufficiently well-behaved function. Integration by parts, recalling that $y_1 \overset{\text{ind}}{\sim} N(\mu_1, 1)$, and that $\phi'(z) = -z\phi(z)$, implies that conditional on $y_{-1} = (y_2, \ldots, y_n)$,

$$\begin{aligned}\mathrm{E}\{(y_1-\mu_1)h_1(y_1,y_{-1})\} &= \int(y_1-\mu_1)\phi(y_1-\mu_1)h_1(y_1,y_{-1})\,\mathrm{d}y_1\\ &= -\left[\phi(y_1-\mu_1)h_1(y_1,y_{-1})\right]_{-\infty}^{\infty}+\int\phi(y_1-\mu_1)h_1'(y_1,y_{-1})\,\mathrm{d}y_1\\ &= \mathrm{E}\left\{h_1'(y_1,y_{-1})\right\},\end{aligned}$$

where $h_1' = \partial h/\partial y_1$, so the same is true unconditionally, and also for indices $2, \ldots, n$.

☐ Now we set

$$h_1(y_1, y_{-1}) = \frac{y_1-\overline{y}}{w} = \frac{y_1-\overline{y}}{\sum_i(y_i-\overline{y})^2}$$

and note that (after a little algebra)

$$h'(y_1,y_{-1}) = \frac{1-n^{-1}}{w} - 2\frac{(y_1-\overline{y})^2}{w^2},$$

which implies that

$$\begin{aligned}\mathrm{E}\left\{\frac{b}{w}\sum_{j=1}^{n}(y_j-\mu_j)(y_j-\overline{y})\right\} &= (n-1)b\mathrm{E}(w^{-1}) - 2\mathrm{E}\left\{\frac{b}{w}\sum_{j=1}^{n}(y_j-\overline{y})^2\right\}\\ &= (n-3)b\mathrm{E}(w^{-1}),\end{aligned}$$

so the central term in (7) has expectation $-2b(n-3)\mathrm{E}(w^{-1})$, from which the result follows directly.

Comments

- ☐ Stein's theorem is the first of a class of results that imply that suitable **shrinkage** (a form of regularisation) can improve frequentist estimation, making the MLE inadmissible in some cases.
- ☐ Examples:
  - the lasso, with shrinkage by soft thresholding;
  - ridge regression, with shrinkage towards the origin;
  - variable selection, with shrinkage by (less stable) hard thresholding.
- ☐ From a theoretical point of view it seems hopeless to try and estimate $O(n)$ parameters, but (we hope that) the effect of shrinkage is that the 'degrees of freedom' p_n is of smaller order than n, i.e., $p_n \to \infty$ as $n \to \infty$ but with $p_n/n \to 0$. If, so, information accumulates about $\mu(\cdot)$, which can be consistently estimated.
- ☐ However, the convergence rate for nonparametric estimation is typically slower than usual parametric rate $O(n^{-1/2})$, and can be as low as $O(n^{-1/5})$.

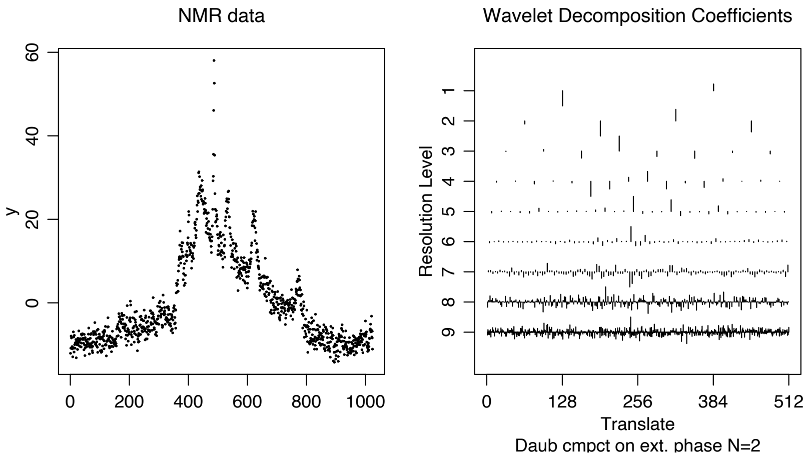
Basis functions

- ☐ We seek to estimate a function $\mu(x)$ based on data $(x_1, y_1), \ldots, (x_n, y_n)$.
- ☐ There are n parameters $\mu_1 = \mu(x_1), \ldots, \mu_n = \mu(x_n)$ (plus noise, …), so we assume that $\mu(x)$ belongs to a suitable class of functions, defined for $x \in \mathcal{X}$.
- ☐ Simple linear model is

$$\mu_{n\times 1} = B_{n\times p}\theta_{p\times 1}, \quad \mathrm{rank}(B) = p \le n,$$

  with the columns of B evaluations at $x_1, \ldots, x_n$ of **basis functions**.
- ☐ The basis functions may be
  - **global** (e.g., polynomials, trigonometric/Fourier functions),
  - **local** (e.g., splines),
  - **multiscale** (e.g., wavelets).
- ☐ We choose the basis for
  - suitability for the problem at hand (e.g., suitably smooth), and
  - computational reasons—want fast, preferably $\mathcal{O}(n)$, handling of $n \times n$ matrices.
- ☐ Later mostly use **spline functions**, on which there is a huge literature.
- ☐ First, **wavelets**:

NMR Data



Left: original data, with $n = 1024$
Right: orthogonal transform with $n = 1024$ coefficients at different resolutions

Orthogonal transformation

- ☐ Original data y with noisy signal $\mu_{n\times 1}$: $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$.
- ☐ Suppose $z_{n\times 1} = W^{\mathrm{T}}_{n\times n} y_{n\times 1}$, where $W^{\mathrm{T}}W = WW^{\mathrm{T}} = I_n$ is orthogonal.
- ☐ Choose W such that $\theta = W^{\mathrm{T}}\mu$ should have many small elements, and a few big ones, giving a sparse representation of μ in the basis corresponding to W;
- ☐ then 'kill' small coefficients of Y, which correspond to the noise, giving

$$\tilde{\theta}_{n\times 1} = \mathrm{kill}(z) = \mathrm{kill}(W^{\mathrm{T}}y);$$

- ☐ then estimate the signal μ by

$$\tilde{\mu} = W\tilde{\theta} = W \times \mathrm{kill}(W^{\mathrm{T}}y),$$

  where the kill$(\cdot)$ operator minimises a suitable risk function.
- ☐ The operator used below applies a form of soft thresholding with parameters chosen from the data.

Wavelet decomposition

- A good choice of W is wavelets, which have nice sparseness properties.
- Here are the Haar wavelet coefficients, with $n = 8$:

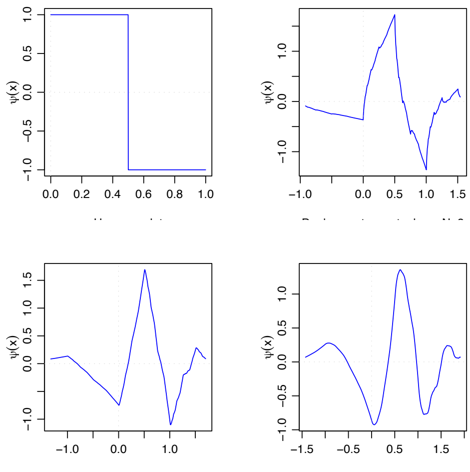
$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

- We set up W so that each column of this orthogonal matrix has unit norm., i.e., we post-multiply this matrix by

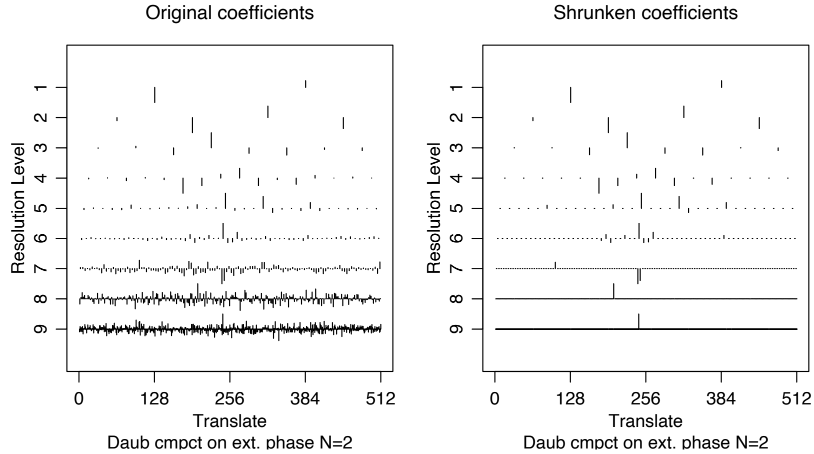
$$\{\text{diag}(\sqrt{8}, \sqrt{8}, 2, 2, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2})\}^{-1},$$

to ensure that $WW^{\mathrm{T}} = W^{\mathrm{T}}W = I_8$.

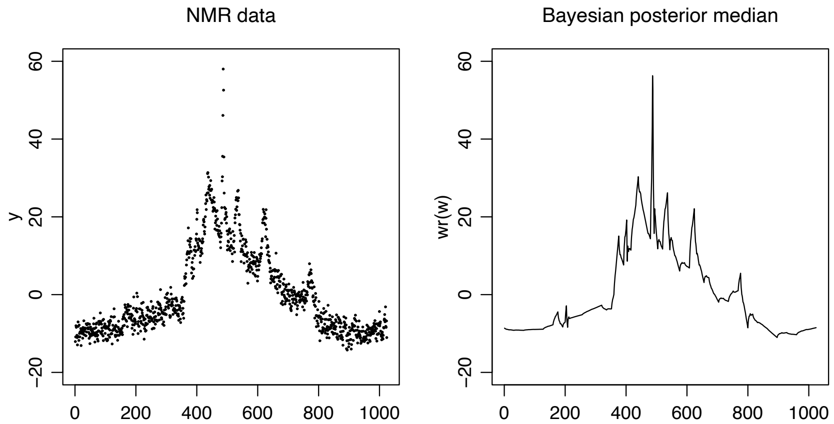
Some Wavelets



NMR data, after transformation and shrinkage



NMR data, after cleaning



Splines

- The **pth degree spline** basis with **knots** $\kappa_1 < \cdots < \kappa_k$ is

$$1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_k)_+^p,$$

  where $u_+ = \max(u, 0)$ is the **positive part function**.
- The resulting matrix B is highly collinear and gives an implausible statistical model.
- **B-spline** bases span the same linear space, but have better numerical properties. They are defined by adding **boundary knots** κ_0 and κ_{k+1} and setting up an **augmented knot sequence**

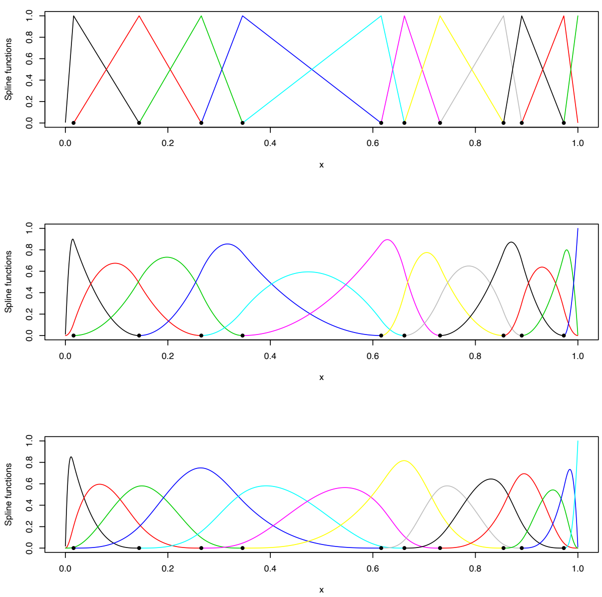
$$\tau_1 \le \cdots \le \tau_M \le \kappa_0 \le \tau_{M+1} = \kappa_1 \le \cdots \le \tau_{M+k} = \kappa_k \le \kappa_{k+1} \le \tau_{k+1+M} \le \cdots \le \tau_{k+2M};$$

  typically the τ_i outside $[\kappa_0, \kappa_{k+1}]$ are set to the boundary knot values. Then

$$\begin{aligned} B_{i,1}(x) &= I(\tau_i \le x < \tau_{i+1}), \quad i = 1, \ldots, k+2M-1, \\ B_{i,m}(x) &= \frac{x-\tau_i}{\tau_{i+m-1}-\tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m}-x}{\tau_{i+m}-\tau_{i+1}} B_{i+1,m-1}(x),\ i = 1, \ldots, k+2M-m, \end{aligned}$$

  where we set $B_{i,1} \equiv 0$ if $\tau_i = \tau_{i+1}$ (avoiding division by zero).
- Cubic splines ($p = 3$, $M = 4$) give visually smooth functions.
- $k = 10$ on the next slide, with $M = 2$ (linear), $M = 3$ (quadratic) and $M = 4$ (cubic).

Linear, quadratic and cubic B-splines



Natural cubic spline

- Suppose the x_j are distinct (no loss of generality, see later) and

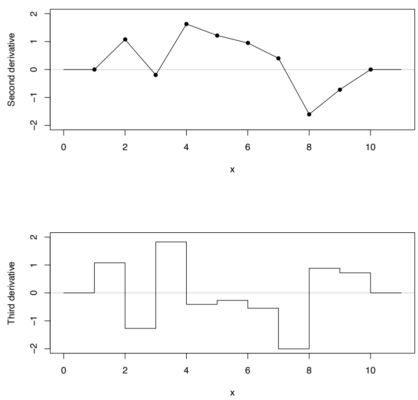
$$a < x_1 < \cdots < x_n < b, \quad \mathcal{X} = [a, b] \subset \mathbb{R}.$$

- A **natural cubic spline** adds the constraint that the function is linear outside $[x_1, x_n]$, and thus avoids high variance due to quadratic and higher terms outside this interval.
- A natural cubic spline
  - has $k = n$ knots, at $x_1 < \cdots < x_n$,
  - is a cubic polynomial on each interval between knots,
  - is continuous, with continuous first and second derivatives at each knot, and
  - is linear on $[a, x_1]$ and $[x_n, b]$, with zero second and third derivatives at x_1 and x_n,
  - has

$$2 + 4(n-1) + 2 \text{ parameters } - 3n \text{ linear constraints } = n$$

    degrees of freedom (df), which can be split into
    - 2 df for a linear fit, plus
    - $n - 2$ df for the second derivatives $\mu''(x_2), \ldots, \mu''(x_{n-1})$.

Natural cubic spline



- A natural cubic spline may constructed by integrating a second derivative function μ'' which is linear and determined by its values at $x = 2, \ldots, 9$; the values at $x = 1, 10$ are zero, and so is μ'' for $x \notin [1, 10]$.
- On integrating twice we gain two constants: $\mu(x) = \beta_0 + \beta_1 x + \int_0^x \int_0^{x'} \mu''(u)\,\mathrm{d}u\,\mathrm{d}x'$.

Optimality of natural cubic splines

- ☐ Let $\mathcal{S}_2(\mathcal{X})$ denote the set of functions μ differentiable on $\mathcal{X} = [a, b]$ with absolutely continuous first derivative μ': there exists an integrable function μ'' such that $\int_a^x \mu''(u)\mathrm{d}u = \mu'(x) - \mu'(a)$ for $x \in \mathcal{X}$.
- ☐ Clearly any μ with two continuous derivatives on $\mathcal{X}$ lies in $\mathcal{S}_2(\mathcal{X})$.

Theorem 23 *Suppose $n \geq 2$, that $a < x_1 < \cdots < x_n < b$, and that μ is the natural cubic spline interpolating the values $y_1, \ldots, y_n$ at $x_1, \ldots, x_n$. If $\tilde{\mu} \in \mathcal{S}_2(\mathcal{X})$ also interpolates the y_j, then*

$$\int_{\mathcal{X}} \tilde{\mu}''^2 \geq \int_{\mathcal{X}} \mu''^2,$$

with equality iff $\tilde{\mu} \equiv \mu$.

- ☐ Thus the natural cubic spline μ minimises the roughness measure $\int_{\mathcal{X}} \mu''^2$ in a larger class of functions than that to which it belongs, making it a natural choice as an interpolant.

Note to Theorem 23

Let $\nu = \tilde{\mu} - \mu \in \mathcal{S}_2(\mathcal{X})$, and note that $\nu(x_j) = 0$ for each j, since $\mu(x_j) = \tilde{\mu}(x_j) = y_j$. The natural boundary conditions imply that $\mu''(a) = \mu''(b) = 0$, so integration by parts yields

$$0 = \left[\mu''(x)\nu'(x)\right]_a^b = \int_{\mathcal{X}} (\mu''\nu')' = \int_{\mathcal{X}} \mu''\nu'' + \int_{\mathcal{X}} \mu'''\nu',$$

and hence the facts that μ''' is piecewise constant and that $\nu(x_j) = 0$ yields

$$\int_{\mathcal{X}} \mu''\nu'' = -\int_{\mathcal{X}} \mu'''\nu' = -\sum_{j=1}^{n-1} \mu'''(x_j^+) \int_{x_j}^{x_{j+1}} \nu' = -\sum_{j=1}^{n-1} \mu'''(x_j^+)\{\nu(x_{j+1}) - \nu(x_j)\} = 0.$$

Hence

$$\int_{\mathcal{X}} \tilde{\mu}''^2 = \int_{\mathcal{X}} (\mu'' + \nu'')^2 = \int_{\mathcal{X}} \mu''^2 + 2\int_{\mathcal{X}} \mu''\nu'' + \int_{\mathcal{X}} \nu''^2 = \int_{\mathcal{X}} \mu''^2 + \int_{\mathcal{X}} \nu''^2 \geq \int_{\mathcal{X}} \mu''^2,$$

wth equality iff $\nu''(x) \equiv 0$. This occurs iff $\nu(x)$ is linear, but since $\nu(x_j) = 0$ at at least two points, $\nu(x) = 0$ for all $x \in \mathcal{X}$.

Roughness penalty

- To choose μ to balance fidelity to the data and smoothness, we chose some $\lambda > 0$ and take $\mu \in \mathcal{S}_2(\mathcal{X})$ to minimise the **penalised sum of squares**

$$\sum_{j=1}^{n}\{y_j - \mu(x_j)\}^2 + \lambda \int_{\mathcal{X}} \mu''(x)^2 \, \mathrm{d}x,$$

  where the second term is a **roughness penalty** on μ'', with a stronger penalty if the **smoothing parameter** $\lambda \to \infty$ and no penalty when $\lambda = 0$.
- Theorem 23 implies that the resulting μ is a natural cubic spline: for any $\tilde{\mu}(x) \in \mathcal{S}_2(\mathcal{X})$ there exists a natural cubic spline $\mu(x)$ such that $\mu(x_j) = \tilde{\mu}(x_j)$, so the sum of squares is the same for $\tilde{\mu}$ and for μ, but

$$\int_{\mathcal{X}} \tilde{\mu}''^2 \geq \int_{\mathcal{X}} \mu''^2,$$

  giving a lower penalty.

Roughness penalty and equivalent formulations

Theorem 24 *In the setting above, the roughness penalty can be expressed as*

$$\int_{\mathcal{X}} \mu''(x)^2 \, \mathrm{d}x = \mu^{\mathrm{T}} K \mu = \gamma^{\mathrm{T}} R \gamma,$$

where

$$\begin{aligned} \mu^{\mathrm{T}} &= (\mu(x_1), \ldots, \mu(x_n)), \quad \gamma^{\mathrm{T}} = (\mu''(x_2), \ldots, \mu''(x_{n-1})), \\ Q^{\mathrm{T}}\mu &= R\gamma, \\ \text{and} \quad & K_{n\times n}, \quad R_{(n-2)\times(n-2)}, \quad Q_{n\times(n-2)} \text{ all have rank } n-2. \end{aligned}$$

Moreover

- *the fit is linear outside (x_1, x_n), so $\mu''(x_1) = \mu''(x_n) = 0$,*
- *Q and R are band matrices, leading to $O(n)$ matrix manipulations, and*
- *imposing the roughness penalty corresponds to using the prior $\gamma \sim \mathcal{N}_{n-2}(0, \sigma^2\psi R^{-1})$ for Bayesian inference, or a mixed effects model*

$$\mu = \beta_0 1_n + \beta_1 x_{n\times 1} + Z_{n\times(n-2)} b, \quad b \sim \mathcal{N}_{n-2}(0, \sigma^2 \psi I_{n-2}),$$

  for a suitable basis matrix Z and $\psi = \lambda^{-1}$.

Note I to Theorem 24

- ☐ Let $\mathcal{X} = [a, b]$ and suppose that $a < x_1 < \cdots < x_n < b$; sometimes below we set $x_0 = a$ and $x_{n+1} = b$. If $\mu(x)$ is a cubic spline on $\mathcal{X}$, then $\mu''(x)$ is continuous and piecewise linear.
- ☐ Taking $\mu''(a) = \mu''(x_1) = \mu''(b) = \mu''(x_n) = 0$ implies that $\mu''(x) = 0$ outside $[x_1, x_n]$, so there is no contribution to $\int(\mu'')^2$ from outside $[x_1, x_n]$; this makes sense because the penalty should not depend on a and b. The optimal $\mu(x)$ will be linear outside $[x_1, x_n]$, because the corresponding μ'' is zero outside this set.
- ☐ Write $\gamma_j = \mu''(x_j)$, parametrise $\mu''(x)$ in terms of $\gamma_2, \ldots, \gamma_{n-1}$, set $h_j = x_j - x_{j-1}$, and obtain

$$\mu''(x) = \frac{\gamma_j(x - x_{j-1}) + \gamma_{j-1}(x_j - x)}{h_j}, \quad x_{j-1} \le x \le x_j, \quad j = 1, \ldots, n+1, \tag{8}$$

where $\gamma_0 = \gamma_1 = \gamma_n = \gamma_{n+1} = 0$; note that $\mu'''(x) = (\gamma_j - \gamma_{j-1})/h_j$ on $[x_{j-1}, x_j]$. Now

$$\begin{aligned}
\int_{x_{j-1}}^{x_j} \mu''(x)^2\,\mathrm{d}x &= \frac{1}{3h_j^2(\gamma_j - \gamma_{j-1})}\left[\gamma_j(x - x_{j-1}) + \gamma_{j-1}(x_j - x)\right]^3\Big|_{x_{j-1}}^{x_j} \\
&= \frac{1}{3h_j^2(\gamma_j - \gamma_{j-1})}\left(h_j^3\gamma_j^3 - h_j^3\gamma_{j-1}^3\right) \\
&= \frac{h_j}{3}(\gamma_j^2 + \gamma_j\gamma_{j-1} + \gamma_{j-1}^2),
\end{aligned}$$

so

$$\int_{\mathcal{X}} \mu''(x)^2\,\mathrm{d}x = \int_{x_1}^{x_n} \mu''(x)^2\,\mathrm{d}x = \sum_{j=2}^{n} \int_{x_{j-1}}^{x_j} \mu''(x)^2\,\mathrm{d}x = \tfrac{1}{3}\sum_{j=2}^{n} h_j(\gamma_j^2 + \gamma_j\gamma_{j-1} + \gamma_{j-1}^2).$$

Note II to Theorem 24

- Recalling that $\gamma_1 = \gamma_n = 0$, we see that $\int_{\mathcal{X}} \mu''(x)^2\,\mathrm{d}x$ equals

$$\frac{1}{3}\left\{h_2\gamma_2^2 + h_3(\gamma_3^2 + \gamma_3\gamma_2 + \gamma_2^2) + \cdots + h_{n-1}(\gamma_{n-1}^2 + \gamma_{n-1}\gamma_{n-2} + \gamma_{n-2}^2) + h_n\gamma_{n-1}^2\right\} = \gamma^{\mathrm{T}} R\gamma,$$

say, where

$$R_{(n-2)\times(n-2)} = \frac{1}{6}\begin{pmatrix} 2(h_2+h_3) & h_3 & 0 & \cdots & 0 \\ h_3 & 2(h_3+h_4) & h_4 & \ddots & \vdots \\ 0 & h_4 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & h_{n-1} \\ 0 & \cdots & 0 & h_{n-1} & 2(h_{n-1}+h_n) \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_2 \\ \vdots \\ \gamma_{n-1} \end{pmatrix}.$$

As $\gamma^{\mathrm{T}} R\gamma$ equals a non-negative integral that is zero iff $\gamma = 0$, R is positive definite, and it is symmetric by construction. It is also strictly diagonal dominant, i.e., $|r_{jj}| > \sum_{i\neq j}|r_{ji}|$ (which also implies that it is invertible).
The unusual labelling of the elements of γ, starting from γ_2, simplifies things below.

- **Aside for time series experts**: if we regard R as the precision (inverse variance) matrix of a Gaussian distribution for γ, then we see that

$$\mathrm{corr}(\gamma_r, \gamma_s \mid \text{ rest }) = \begin{cases} -h_{r+1}/\{4(h_r+h_{r+1})(h_{r+1}+h_{r+2})\}^{1/2}, & s = r+1, \\ 0, & \text{otherwise}, \end{cases}$$

corresponding to a Markov process. If the x_j are equally-spaced, then all the h_j are equal and $\mathrm{corr}(\gamma_r, \gamma_s \mid \text{ rest }) = I(r = s) - I(|r-s| = 1)/4$ is PACF of an AR(1) model.

- Integration by parts yields

$$\int_{\mathcal{X}} \mu''(x)^2\,\mathrm{d}x = \left[\mu'(x)\mu''(x)\right]_a^b - \int_{\mathcal{X}} \mu'''(x)\mu'(x)\,\mathrm{d}x,$$

where the first term on the right-hand side equals zero because $\mu''(a) = \mu''(b) = 0$, and as the third derivative is constant on each interval $(x_1, x_2], \ldots, (x_{n-1}, x_n]$ and zero elsewhere, the second term on the right-hand side equals

$$\begin{aligned} -\sum_{j=1}^{n-1}\mu'''(x_j^+)\int_{x_j}^{x_{j+1}}\mu'(x)\,\mathrm{d}x &= \sum_{j=1}^{n-1}\frac{\gamma_{j+1}-\gamma_j}{h_{j+1}}(\mu_j - \mu_{j+1}) \\ &= \sum_{j=2}^{n-1}\gamma_j\left(\frac{\mu_{j+1}-\mu_j}{h_{j+1}} - \frac{\mu_j - \mu_{j-1}}{h_j}\right) \\ &= \gamma^{\mathrm{T}} Q^{\mathrm{T}}_{(n-2)\times n}\mu, \end{aligned} \qquad (9)$$

say, because $\gamma_1 = \gamma_n = 0$; here and below, $\mu^{\mathrm{T}} = (\mu_1, \ldots, \mu_n)$, where $\mu_j = \mu(x_j)$. If we label the $n-2$ rows of $Q^{\mathrm{T}}\mu$ like those of γ (i.e., starting at row $j = 2$), then the jth row of $Q^{\mathrm{T}}\mu$ is

$$\frac{\mu_{j+1}-\mu_j}{h_{j+1}} - \frac{\mu_j - \mu_{j-1}}{h_j}, \quad j = 2, \ldots, n-1,$$

so Q^{T} is tridiagonal and fast (i.e., linear in n) matrix operations with it are possible.

- Note that $Q^{\mathrm{T}} 1_n = 0$ and $Q^{\mathrm{T}} x = 0$, which verifies that linear functions are not penalised.

Note III to Theorem 24

- ☐ If we integrate $\mu''(x)$ twice, we see that for $x_{j-1} \le x \le x_j$ we can write

$$\begin{aligned}\mu(x) &= \frac{(x-x_{j-1})\mu_j + (x_j - x)\mu_{j-1}}{h_j} \\ &\quad -\tfrac{1}{6}(x-x_{j-1})(x_j-x)\left\{\left(1+\frac{x-x_{j-1}}{h_j}\right)\gamma_{j-1} + \left(1+\frac{x_j-x}{h_j}\right)\gamma_j\right\}.\end{aligned}$$

Hence the first derivatives at $x_1, \ldots, x_n$ from the right and from the left may be written as

$$\begin{aligned}\mu'(x_j^+) &= (\mu_{j+1}-\mu_j)/h_{j+1} - \tfrac{1}{6}h_{j+1}(2\gamma_j + \gamma_{j+1}), \\ \mu'(x_j^-) &= (\mu_j - \mu_{j-1})/h_j + \tfrac{1}{6}h_j(\gamma_{j-1}+2\gamma_j).\end{aligned}$$

As $\gamma_1 = \gamma_n = 0$ the limits at x_1 and x_n reduce to

$$\mu'(x_1^+) = \frac{\mu_2-\mu_1}{h_2} - \tfrac{1}{6}h_2\gamma_2, \quad \mu'(x_n^-) = \frac{\mu_n - \mu_{n-1}}{h_n} + \tfrac{1}{6}h_n\gamma_{n-1}. \tag{10}$$

Outside $[x_1, x_n]$ the function $\mu(x)$ is linear and may be written as

$$\mu(x) = \begin{cases} \mu_1 - (x_1 - x)\mu'(x_1^+), & x \le x_1, \\ \mu_n + (x-x_n)\mu'(x_n^-), & x \ge x_n, \end{cases}$$

which guarantees that $\mu'(x)$ is continuously differentiable at x_1 and x_n.

- ☐ Now $\mu(x)$ is everywhere continuous, has $\mu''(x_j^+) = \mu''(x_j^-) = \gamma_j$ for each j, and μ' is continuous at x_1 and x_n, so to make μ a natural cubic spline we must ensure that μ' is continuous at $x_2, \ldots, x_{n-1}$, i.e., $\mu'(x_j^-) = \mu'(x_j^+)$ for $j = 2, \ldots, n-1$. Thus (10) yields

$$\frac{\mu_{j+1}-\mu_j}{h_{j+1}} - \frac{\mu_j - \mu_{j-1}}{h_j} = \tfrac{1}{6}h_j\gamma_{j-1} + \tfrac{1}{3}(h_j + h_{j+1})\gamma_j + \tfrac{1}{6}h_{j+1}\gamma_{j+1}, \quad j = 2, \ldots, n-1,$$

which on comparison with (9) can be written in matrix form as $Q^{\mathrm{T}}\mu = R\gamma$.

- ☐ Hence using invertibility of R we can write the penalty as

$$\int_{\mathcal{X}} \mu''(x)^2 \, \mathrm{d}x = \gamma^{\mathrm{T}}Q^{\mathrm{T}}\mu = \gamma^{\mathrm{T}}R\gamma = \mu^{\mathrm{T}}QR^{-1}Q^{\mathrm{T}}\mu = \mu^{\mathrm{T}}K\mu.$$

- ☐ The $n \times 1$ vector μ lies in the vector space $\mathbb{R}^n$, and we can form a basis of this space starting with the vectors 1_n and $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ and then adding any $n-2$ linearly independent vectors. In particular, on integrating (8) up to x, twice, we see that we can write

$$\mu(x) = \beta_0 + \beta_1 x + \sum_{j=2}^{n-1} p_j(x)\gamma_j, \quad x \in \mathcal{X}, \tag{11}$$

where the term $\beta_0 + \beta_1 x$ comes from the constants of integration and the cubic polynomials $p_j(x)$ stem from integrating the linear functions $\mu''(x)$ in each of the intervals $[x_{j-1}, x_j]$. On evaluating (11) at $x_1, \ldots, x_n$ we obtain

$$\mu = \beta_0 1_n + \beta_1 x + P\gamma, \quad \beta_0, \beta_1, \gamma_2, \ldots \gamma_{n-1} \in \mathbb{R},$$

where the $n \times (n-2)$ matrix P has rank $n-2$, because its first row consists of zeros and rows $2, \ldots, n-1$ are lower triangular (as $\mu(x)$ involved integration from a to x).

Note IV to Theorem 24

- ☐ Recall that $Q^{\mathrm{T}}1_n = Q^{\mathrm{T}}x = 0$, giving $Q^{\mathrm{T}}\mu = Q^{\mathrm{T}}P\gamma$, and thus

$$\gamma^{\mathrm{T}}R\gamma = \mu^{\mathrm{T}}QR^{-1}Q^{\mathrm{T}}\mu = \gamma^{\mathrm{T}}P^{\mathrm{T}}QR^{-1}Q^{\mathrm{T}}P\gamma = \gamma^{\mathrm{T}}(P^{\mathrm{T}}QR^{-1})R(R^{-1}Q^{\mathrm{T}}P)\gamma,$$

which implies that $P^{\mathrm{T}}QR^{-1} = I_{n-2}$. Hence $Q_{n\times(n-2)}$ and $K_{n\times n} = QR^{-1}Q^{\mathrm{T}}$ have rank $n-2$.

- ☐ Now consider a model in which $\mu = \beta_0 + \beta_1 x + P\gamma$ and γ is random, with

$$y \mid \gamma \sim \mathcal{N}_n(\mu, \sigma^2 I_n), \quad \gamma \sim \mathcal{N}_{n-2}(0, \sigma^2\lambda^{-1}R^{-1}), \quad \lambda > 0. \tag{12}$$

Apart from constants, the logarithm of the posterior density $f(\beta_0, \beta_1, \gamma \mid y; \sigma^2, \lambda)$ equals

$$\log f(y \mid \gamma) + \log f(\gamma) = -\left\{(y-\mu)^{\mathrm{T}}(y-\mu) + \lambda\gamma^{\mathrm{T}}R\gamma\right\}/(2\sigma^2) - \frac{2n-2}{2}\log\sigma^2 + \frac{n-2}{2}\log\lambda,$$

and for fixed σ^2 and λ the values of β_0, β_1 and γ that maximise $f(\beta_0, \beta_1, \gamma \mid y; \sigma^2, \lambda)$ are found by minimising the penalised sum of squares

$$(y-\mu)^{\mathrm{T}}(y-\mu) + \lambda\gamma^{\mathrm{T}}R\gamma = \sum_{j=1}^{n}\{y_j - \mu(x_j)\} + \lambda\int_{\mathcal{X}} \mu''(x)^2\,\mathrm{d}x.$$

- ☐ As R is not diagonal, the elements of γ in (12) are not independent, but we can use the spectral decomposition to diagonalise R, writing $\gamma^T R\gamma = (A\gamma)^{\mathrm{T}}(A\gamma) = b^{\mathrm{T}}b$, say, where A is invertible and $R = A^{\mathrm{T}}A$, so $R^{-1} = A^{-1}(A^{\mathrm{T}})^{-1}$ and $AR^{-1}A^{\mathrm{T}} = I_{n-2}$. Thus

$$b = A\gamma \sim \mathcal{N}_{n-2}(0, \sigma^2\lambda^{-1}AR^{-1}A^{\mathrm{T}}) = \mathcal{N}_{n-2}(0, \sigma^2\psi I_{n-2}), \quad \psi = \lambda^{-1}.$$

In this parametrization we have

$$\mu = \beta_0 1_n + \beta_1 x + P\gamma = \beta_0 1_n + \beta_1 x + Zb, \quad Z = PA^{-1},$$

and with the stated normal distribution for b we see that this is a mixed model with fixed effect $\beta_0 1_n + \beta_1 x$ and random effect Zb.

- ☐ Hence imposing the ‘natural’ smoothness penalty based on $\int(\mu'')^2$ is equivalent to Bayesian inference using an improper (constant) prior for β_0 and β_1 and a proper Gaussian prior for γ,

$$\gamma \sim \mathcal{N}_{n-2}(0, \sigma^2\psi R^{-1})$$

or using the mixed effects model with

$$\mu = X\beta + Zb = \beta_0 1_n + \beta_1 x + Zb, \quad b \sim \mathcal{N}_{n-2}(0, \sigma^2\psi I_{n-2}),$$

or any other equivalent formulation in terms of an invertible transformation of the columns of Z.

- ☐ In this setup we could set $\psi = 0$ (equivalent to $\lambda = \infty$), corresponding to fitting the fixed effects alone. Testing $\psi = 0$ against $\psi > 0$ corresponds to testing a linear model against the spline fit in which $b \neq 0$.

Penalised fitting

- On writing

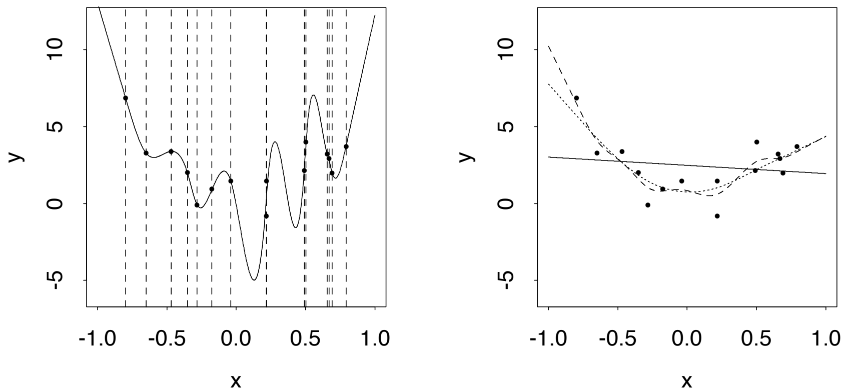
$$\sum_{j=1}^{n}\{y_j-\mu(x_j)\}^2+\lambda\int_{\mathcal{X}}\mu''(x)^2\,\mathrm{d}x \;=\; (y-\mu)^{\mathrm{T}}(y-\mu)+\lambda\mu^{\mathrm{T}}K\mu$$
$$= \mu^{\mathrm{T}}(I_n+\lambda K)\mu-2y^{\mathrm{T}}\mu+y^{\mathrm{T}}y,$$

  we see that the fitted values are

$$\widehat{\mu}_{n\times 1}=(I_n+\lambda K)^{-1}y_{n\times 1}=S_\lambda y$$

  in terms of the **smoothing matrix** S_λ.
- The natural cubic spline will change as λ varies:
  - as $\lambda\to 0$ (equivalently, $\psi\to\infty$), we have $S_\lambda\to I_n$ and we obtain the spline interpolating $y_1,\ldots,y_n$, which has n df, corresponding to the n elements of μ;
  - as $\lambda\to\infty$ (equivalently, $\psi\to 0$), we recover the best-fitting straight line $\mu(x)=\beta_0+\beta_1 x$, for which $\mu''(x)=0$, with 2 df; and
  - intermediate values of λ correspond to intermediate fits.

Natural cubic



- The unpenalized natural cubic spline on the left passes perfectly through the 15 points, but is very sensitive to their values: think of beads on wires.
- On the right we see the effect of a roughness penalty with 2 (solid), 7 (dashes) and 3.7 (dots) df.

Variants

- Obvious generalisation allows weight matrix $W = \mathrm{diag}(w_1, \ldots, w_n)$.
- If the $x_1, \ldots, x_n$ are not unique, write $\mathrm{E}(y) = N_{n\times n'}\mu_{n'\times 1}$ in terms of the means μ at the n' unique elements of x, and minimise

$$(y - N\mu)^{\mathrm{T}} W (y - N\mu) + \lambda\mu^{\mathrm{T}} K\mu.$$

  where $K_{n'\times n'}$ arises as before as the roughness penalty on $\mu(x)$.
- Could use a truncated power basis of degree $p-1$ with knots at $\kappa_1, \ldots, \kappa_q$, i.e.,

$$\mu(x) = \beta_0 + \beta_1 x + \cdots + \beta_{p-1}x^{p-1} + \sum_{i=1}^{q} b_i (x - \kappa_i)_+^{p-1},$$

  set $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, b^{\mathrm{T}})$ where $\beta^{\mathrm{T}} = (\beta_0, \ldots, \beta_{p-1})$, $b^{\mathrm{T}} = (b_1, \ldots, b_q)$ and minimise

$$(y - B\theta)^{\mathrm{T}} W (y - B\theta) + \lambda\theta^{\mathrm{T}} D\theta$$

  where the penalty matrix D does not affect β (e.g., $D = \mathrm{diag}(0, \ldots, 0, I_q)$). This gives

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}}WB + \lambda D)^{-1}B^{\mathrm{T}}Wy, \quad \widehat{y} = B(B^{\mathrm{T}}WB + \lambda D)^{-1}B^{\mathrm{T}}Wy = S_\lambda y.$$

  Since β is unpenalised, $S_\lambda y$ tends to the polynomial fit when $\lambda \to \infty$.

Modelling choices

- Choose θ to minimise

$$\sum_{j=1}^{n} \{y_j - \theta^{\mathrm{T}} B(x_j)\}^2 + \lambda\theta^{\mathrm{T}} D\theta$$

  for basis functions $B(\cdot)$, D symmetric positive semi-definite and some $\lambda > 0$.
- Primary choices (affect fit):
  - spline model, i.e., degree, knot locations, boundary constraints (if any);
  - form of the penalty (imposes smoothness through prior variance structure for θ).
- Simple choice of number of knots is $q = \min(35, n_u/4)$, where n_u is the number of unique x_j, placed at $(i+1)/(q+2)$ quantiles of the unique x_j, for $i = 1, \ldots, q$.
- Secondary choices (only affect fit through any numerical error):
  - basis functions, chosen for interpretability;
  - basis functions as used in the computations—by setting $B = B_* L^{-1}$, where $L_{p\times p}$ is invertible, we can transform any choice of basis functions B to the (very stable and numerically efficient) B-spline basis B_* and then set

$$\widehat{y} = B(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}y = B_*(B_*^{\mathrm{T}}B_* + \lambda L^{\mathrm{T}}DL)^{-1}B_*^{\mathrm{T}}y,$$

    so no need to worry about numerical aspects.

Choice of λ by cross-validation

- Fitted values are $\widehat{y} = S_\lambda y$.
- Fitted value $\widehat{y}_{-j}$ for y_j obtained when (x_j, y_j) is dropped from fit is given by

$$S_{jj}(\lambda)(y_j - \widehat{y}_{-j}) = \widehat{y}_j - \widehat{y}_{-j}.$$

- Cross-validation sum of squares is

$$\mathrm{CV}(\lambda) = \sum_{j=1}^{n}(y_j - \widehat{y}_{-j})^2 = \sum_{j=1}^{n}\left\{\frac{y_j - \widehat{y}_j}{1 - S_{jj}(\lambda)}\right\}^2,$$

  and generalised cross-validation sum of squares is

$$\mathrm{GCV}(\lambda) = \sum_{j=1}^{n}\left\{\frac{y_j - \widehat{y}_j}{1 - \mathrm{tr}(S_\lambda)/n}\right\}^2,$$

  where $S_{jj}(\lambda)$ is (j,j) element of S_λ.
- Both tend to give curves that are too variable, and so do approaches based on minimising Mallows' C_p.
- Better to use mixed model formulation and REML, if normal model credible . . .

Mixed model: Reminder

- Linear mixed model formulation

$$y \mid b \;\sim\; \mathcal{N}_n(X\beta + Zb, \Omega), \quad b \;\sim\; \mathcal{N}_q(0, \Omega_b),$$

  yields

$$y \sim \mathcal{N}_n(X\beta, Z\Omega_b Z^{\mathrm{T}} + \Omega), \quad Z\Omega_b Z^{\mathrm{T}} + \Omega = \sigma^2\Delta^{-1}(\psi),$$

  with ψ the vector of distinct variance ratios and log likelihood

$$\ell(\beta, \sigma^2, \psi) \equiv -\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}\Delta(y - X\beta) - \frac{n}{2}\log\sigma^2 + \tfrac{1}{2}\log|\Delta|.$$

- REML inference for ψ, σ^2 is based on the restricted likelihood

$$\ell_{\mathrm{R}}(\psi, \sigma^2) \equiv \tfrac{1}{2}\log|\Delta| - \tfrac{1}{2}\log|X^{\mathrm{T}}\Delta X| - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Delta(y - X\widehat{\beta}_\psi) - \frac{n-p}{2}\log\sigma^2,$$

  where

$$\widehat{\beta}_\psi = (X^{\mathrm{T}}\Delta X)^{-1}X^{\mathrm{T}}\Delta y,$$

  which involves taking a grid of values of ψ, or iterating on ψ, or . . .

(RE)ML estimation of λ

- ☐ With $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, b^{\mathrm{T}})$, the penalised sum of squares

$$\sum_{j=1}^{n}\{y_j - \theta^{\mathrm{T}} B(x_j)\}^2 + \lambda\theta^{\mathrm{T}} D\theta = (y - B\theta)^{\mathrm{T}}(y - B\theta) + \lambda\theta^{\mathrm{T}} D\theta$$

  is obtained as $-2\sigma^2$ times the likelihood exponent on setting

$$y \mid b \sim \mathcal{N}_n(X\beta + Zb, \sigma^2 I_n), \quad b \sim \mathcal{N}_q(0, \sigma_b^2 I_q),$$

  where
  - $B = (X, Z)$ is an $n \times (p + q)$ matrix,
  - $X_{n\times p}$ corresponds to the unpenalized columns of B,
  - $Z_{n\times q}$ corresponds to the (transformed?) penalized columns of B,
  - $D = \mathrm{diag}(0_p, 1_q)$, and
  - $\sigma_b^2 = \sigma^2/\lambda$, so in the general notation, $\psi = 1/\lambda$ and $\Delta(\psi) = (I_n + \lambda^{-1} ZZ^{\mathrm{T}})^{-1}$.
- ☐ Can estimate σ^2 and λ using ML or REML (better), giving $\widehat{\lambda} = \widehat{\sigma}^2/\widehat{\sigma}_b^2$.
- ☐ Comparison with GCV on page 106 of notes.

A useful lemma

Lemma 25 *Let A and B be $q \times q$ positive semi-definite matrices, and suppose that $(A + \lambda B)^{-1}$ exists for some $\lambda > 0$. Let η be an eigenvalue of $(A + \lambda B)^{-1} A$. Then*

- ☐ *if B is invertible, then*

$$\eta = \frac{\eta'}{\lambda + \eta'},$$

  where η' is an eigenvalue of $B^{-1/2} A B^{-1/2}$;
- ☐ *if A is invertible, then*

$$\eta = \frac{1}{1 + \lambda\eta''},$$

  where η'' is an eigenvalue of $A^{-1/2} B A^{-1/2}$;

Note to Lemma 25

Exercise!

Equivalent degrees of freedom

- Least squares estimation gives $\widehat{y} = Hy$, with $\mathrm{tr}(H) = p$, in terms of the projection matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.
- In the spline case $X_{n\times p}$ is replaced by $B_{n\times(p+q)} = (X, Z)$.
- Define the **equivalent degrees of freedom** as $\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda)$, i.e.,

$$\mathrm{df}_\lambda = \mathrm{tr}\{B(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}\} = \mathrm{tr}\{(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}B\} = \sum_{j=1}^{p+q} \frac{1}{1+\eta_j\lambda}$$

  where $\eta_1, \ldots, \eta_{p+q} \in [0,1]$ are eigenvalues of $(B^{\mathrm{T}}B)^{-1/2}D(B^{\mathrm{T}}B)^{-1/2}$.
- Since perfectly polynomial data of degree $p-1$ would be unchanged by smoothing, p of the eigenvalues of S_λ must equal 1: since $D_{(p+q)\times(p+q)}$ has rank q, p of the η_j equal 0.
- Clearly $\mathrm{tr}(S_\lambda)$ is monotone decreasing in λ, with

$$\mathrm{tr}(S_\lambda) \to \begin{cases} p, & \lambda \to \infty, \\ p+q, & \lambda \to 0. \end{cases}$$

- Hence we can specify smoothness using either λ or df_λ (easier to interpret).

Note on equivalent degrees of freedom

Using the lemma we obtain

$$\mathrm{tr}(S_\lambda) = \mathrm{tr}\left\{B(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}\right\} = \mathrm{tr}\left\{(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}B\right\} = \sum_{j=1}^{p+q} \frac{1}{1+\lambda\eta_j},$$

where $B^{\mathrm{T}}B$ is invertible and $\eta_1 < \cdots < \eta_{p+q}$ are the eigenvalues of $(B^{\mathrm{T}}B)^{-1/2}D(B^{\mathrm{T}}B)^{-1/2}$. Since D has rank q, $\eta_1 = \cdots = \eta_p = 0$, and therefore

$$p \le \sum_{j=1}^{p+q} \frac{1}{1+\lambda\eta_j} \le p+q,$$

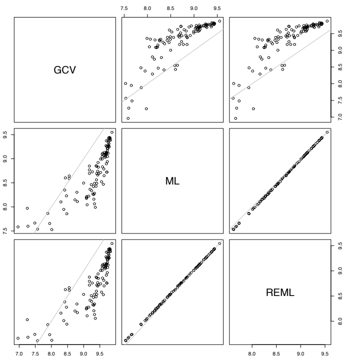
with the limits attained when $\lambda \to \infty$ and $\lambda \to 0$ respectively.

Simulation

Equivalent degrees of freedom (EDF) estimated by GCV, ML and REML for 100 replicate datasets simulated by

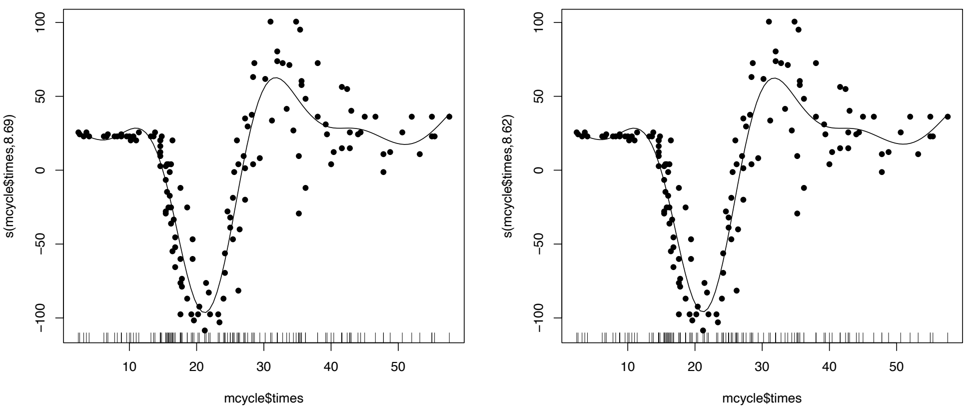
$$y_j \mid x_j \overset{\text{ind}}{\sim} \mathcal{N}\left\{1.5\phi\left(\frac{x_j - 0.35}{0.15}\right) - \phi\left(\frac{x_j - 0.8}{0.04}\right), 0.1^2\right\}, \quad x_j \overset{\text{iid}}{\sim} U(0,1), \quad j = 1, \dots, 100.$$

The EDF for GCV is almost always larger than for the other two, which are essentially identical in this case.



Motorcycle data

Splines fits estimated using GCV (left) and REML (right), giving EDF=8.69 and 8.62. Both fits are dodgy for small x because of the heteroscedasticity.



Comments

- ☐ Spline smoothing is widely used because:
  - it is flexible, links nicely to parametric regression and mixed models;
  - automatic choices of number and placing of knots feasible;
  - under normality assumptions, the smoothing parameter can be chosen in a principled way using ML or REML, with a Bayesian interpretation;
  - can easily remove the assumption of unique x_j, only made for convenience;
  - extends to several smooth functions, or to higher dimensions, in a natural way, using appropriate basis functions and penalties.
- ☐ Extends to other bases (support vector machines, Gaussian processes, ...)
- ☐ Links naturally to Bayesian modelling, by setting joint prior density on β, Ω, Ω_b, number and placing of knots, ...
- ☐ Next: inference ...

Inference for Spline Fits slide 179

Error of a linear smoother

Lemma 26 *A linear smoother giving $\widehat{y} = Sy$ applied to data $y \sim (\mu, \sigma^2 I_n)$ satisfies*

$$\mathrm{E}\left\{\sum_{j=1}^{n}(\widehat{y}_j - \mu_j)^2\right\} = \|(I-S)\mu\|^2 + \sigma^2 \mathrm{tr}(S^{\mathrm{T}}S),$$

where terms on the right represent the squared bias and variance contributions to the overall mean squared error.

This suggests that (as in the usual linear model):

- ☐ there is a bias $(I - S)\mu$ unless μ lies in the kernel of the matrix $I - S$, which will happen only if
  - the columns of the fitted X contain all the true covariates, and
  - there is no need for smoothing, i.e., the spline columns are unnecessary;
- ☐ another EDF definition could be $\mathrm{tr}(S^{\mathrm{T}}S)$, since in the linear model case, the second term would be $\mathrm{tr}(H^{\mathrm{T}}H) = \mathrm{tr}(H) = p$.

Note to Lemma 26

Since

$$(\widehat{y}_j - \mu_j)^2 = \{\widehat{y}_j - \mathrm{E}(\widehat{y}_j)\}^2 + 2\{\widehat{y}_j - \mathrm{E}(\widehat{y}_j)\}\{\mathrm{E}(\widehat{y}_j) - \mu_j\} + \{\mathrm{E}(\widehat{y}_j) - \mu_j\}^2,$$

the expectation is clearly

$$\begin{aligned}
\sum_{j=1}^{n} \mathrm{var}\,(\widehat{y}_j) + \sum_{j=1}^{n} \{\mathrm{E}(\widehat{y}_j) - \mu_j\}^2 &= \mathrm{tr}\,\{\mathrm{var}(\widehat{y})\} + (S\mu - \mu)^{\mathrm{T}}(S\mu - \mu) \\
&= \mathrm{tr}\,\{\mathrm{var}(Sy)\} + \|(I - S)\mu\|^2 \\
&= \mathrm{tr}\{S\mathrm{var}(y)S^{\mathrm{T}}\} + \|(I - S)\mu\|^2 \\
&= \sigma^2 \mathrm{tr}(SS^{\mathrm{T}}) + \|(I - S)\mu\|^2,
\end{aligned}$$

as required.

Inference

- So far we have discussed only 'point estimation' of the smooth function $\mu(x)$, but in applications we also want
  - pointwise confidence intervals for $\mu(x)$,
  - overall confidence bands for $\{\mu(x) : x \in \mathcal{S}\}$, where $\mathcal{S}$ is some subset of $\mathcal{X}$, and
  - tests of hypotheses such as 'is the spline part needed?' and 'is the curve monotonic?'
- Simplest approach to pointwise inference for $\mu(x)$ is to note that since the spline estimator is linear, we can write $\widehat{\mu}(x) = a_x^{\mathrm{T}} y$, giving

$$\mathrm{var}\,\{\widehat{\mu}(x)\} = a_x^{\mathrm{T}} \mathrm{var}(y) a_x = \sigma^2 a_x^{\mathrm{T}} a_x$$

  when $\mathrm{var}(y) = \sigma^2 I_n$.
- This leads to the $(1 - 2\alpha)$ **variability band**,

$$\widehat{\mu}(x) \pm \widehat{\sigma}\sqrt{a_x^{\mathrm{T}} a_x} \times t_{\mathrm{df}_{\mathrm{res}}}(1 - \alpha),$$

  where $\mathrm{df}_{\mathrm{res}}$ is the residual degrees of freedom; this does not account for the bias of $\widehat{\mu}(x)$ or the estimation of λ.
- The use of $t_{\mathrm{df}_{\mathrm{res}}}(1 - \alpha)$ is intended to allow for estimation of σ; for a pointwise 95% variability band we often take $t_{\mathrm{df}_{\mathrm{res}}}(1 - \alpha) \approx 2$.

Mixed model

☐ Recall that in the linear mixed model

$$y \mid b \sim \mathcal{N}_n(X\beta + Zb, \Omega), \quad b \sim \mathcal{N}_q(0, \Omega_b),$$

we have predictors/estimates (both with circumflexes and called BLUPs for short)

$$\widehat{\beta} = (X^{\mathrm{T}}\Delta X)^{-1}X\Delta y, \quad \widehat{b} = \Omega_b Z^{\mathrm{T}}\Delta(y - X\widehat{\beta}),$$

where $\Delta = \sigma^2(\Omega + Z\Omega_b Z^{\mathrm{T}})^{-1}$ is evaluated at the (RE)ML estimates $\widehat{\sigma}^2$, $\widehat{\psi}$.

☐ In our case we write

$$y \mid b \sim \mathcal{N}_n(B\theta, \sigma^2 I_n), \quad b \sim \mathcal{N}_q(0, \sigma^2\lambda^{-1}I_q),$$

where $B = (X, Z)$ is the $n \times (p+q)$ basis matrix, and

– $\theta = \begin{pmatrix}\beta \\ b\end{pmatrix}$ represents the linear unknowns,

– $X_{n\times p}$ corresponds to the (unshrunk) polynomial columns of the basis matrix B,

– $Z_{n\times q}$ corresponds to the spline columns of B.

☐ The BLUP $\widehat{\mu}(x) = B_x\widehat{\theta} = X_x\widehat{\beta} + Z_x\widehat{b}$ of $\mu(x)$ predicts $B_x\theta = X_x\beta + Z_x b$, where

$$B_x = (X_x, Z_x), \quad X_x = (1, x, \ldots, x^{p-1}), \quad Z_x = ((x-\kappa_1)_+^{p-1}, \ldots, (x-\kappa_q)_+^{p-1}).$$

Properties of BLUPs

☐ In the mixed model set-up, the target quantity $\mu(x) = B_x^{\mathrm{T}}\theta$ is regarded as random, and we have two bases for inference:

– compute uncertainties conditional on the realised value of b;

– treat b as random and use unconditional uncertainty computations.

☐ The following lemma gives the conditional and unconditional means and variances.

Lemma 27 *If in the mixed linear model $y = X\beta + Zb + \varepsilon = B\theta + \varepsilon$ we write*

$$\theta = \begin{pmatrix}\beta \\ b\end{pmatrix}, \quad \widehat{\theta} = \begin{pmatrix}\widehat{\beta} \\ \widehat{b}\end{pmatrix},$$

then $\widehat{\theta} - \theta = (B^{\mathrm{T}}B + \lambda D)^{-1}(B^{\mathrm{T}}\varepsilon - \lambda D\theta)$ satisfies

$$\begin{aligned}
\mathrm{E}(\widehat{\theta} - \theta \mid b) &= -\lambda(B^{\mathrm{T}}B + \lambda D)^{-1}\begin{pmatrix}0 \\ b\end{pmatrix}, \\
\mathrm{E}(\widehat{\theta} - \theta) &= 0, \\
\mathrm{var}(\widehat{\theta} - \theta \mid b) &= \sigma^2(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}B(B^{\mathrm{T}}B + \lambda D)^{-1}, \\
\mathrm{var}(\widehat{\theta} - \theta) &= \sigma^2(B^{\mathrm{T}}B + \lambda D)^{-1}.
\end{aligned}$$

Note to Lemma 27

- ☐ Conditional on b, we have $y \sim \mathcal{N}_n(X\beta + Zb, \sigma^2 I_n)$, so $y \sim \mathcal{N}_n(B\theta, \sigma^2 I_n)$. In this case $\widehat{\theta}$ is just the value that minimises
$$(y - B\theta)^{\mathrm{T}}(y - B\theta) + \lambda\theta^{\mathrm{T}} D\theta,$$
with $D = \mathrm{diag}(0_p, 1_q)$, i.e., $\widehat{\theta} = (B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}y = AB^{\mathrm{T}}y$, say.
- ☐ For the first result, we have
$$\widehat{\theta} - \theta = AB^{\mathrm{T}}y - \theta = AB^{\mathrm{T}}(B\theta + \varepsilon) - A(B^{\mathrm{T}}B + \lambda D)\theta = AB^{\mathrm{T}}\varepsilon - \lambda AD\theta$$
which has expectation $-\lambda AD\theta$, which is the stated formula, because $D\theta = (0^{\mathrm{T}}, b^{\mathrm{T}})^{\mathrm{T}}$.
- ☐ The second result is immediate from the first, as $\mathrm{E}(b) = 0$.
- ☐ The third result is also immediate from the first, as $\mathrm{var}(\varepsilon) = \sigma^2 I_n$ and $\mathrm{var}(\theta \mid b) = 0$, giving $\mathrm{var}(\widehat{\theta} \mid b) = \sigma^2 AB^{\mathrm{T}}BA^{\mathrm{T}}$, as stated.
- ☐ For the final result, note that here θ is random, independent of ε, and hence
$$\mathrm{var}(\widehat{\theta} - \theta) = A\mathrm{var}(B^{\mathrm{T}}\varepsilon - \lambda D\theta)A^{\mathrm{T}},$$
with the inner variance becoming
$$B^{\mathrm{T}}\sigma^2 I_n B + (\lambda)^2 D^2\sigma^2/\lambda = \sigma^2(B^{\mathrm{T}}B + \lambda D) = \sigma^2 A^{-1},$$
because $\mathrm{var}(b) = \sigma^2\lambda^{-1}I_q$, giving $\mathrm{var}(\widehat{\theta} - \theta) = A\sigma^2 A^{-1}A = \sigma^2 A^{-1}$, which equals the stated formula.

Conditional analysis

- ☐ Lemma 27 gives us the variance and mean of the normally-distributed quantity
$$\widehat{\mu}(x) = B_x\widehat{\theta}$$
and it follows that, conditionally on b,
$$\left.\frac{\widehat{\mu}(x) - \mathrm{E}\{\widehat{\mu}(x)\}}{\sqrt{\mathrm{var}\{\widehat{\mu}(x)\}}}\ \right|\ b \quad \sim \quad \mathcal{N}(0, 1),$$
where
$$\begin{aligned}\mathrm{E}\{\widehat{\mu}(x) \mid b\} &= B_x(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}B\theta,\\ \mathrm{var}\{\widehat{\mu}(x) \mid b\} &= \sigma^2 B_x(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}B(B^{\mathrm{T}}B + \lambda D)^{-1}B_x^{\mathrm{T}},\end{aligned}$$
which yields approximate confidence intervals for $\mathrm{E}\{\widehat{\mu}(x)\}$ when the parameters are replaced by estimates.
- ☐ If the bias is small, then this interval is also an interval for $\mu(x) \approx \mathrm{E}\{\widehat{\mu}(x)\}$.
- ☐ But this interval will fail if the bias is large, i.e., when μ'' is not small . . . it needs to be wider . . .

Unconditional analysis

☐ Lemma 27 implies that the unconditional bias of $\widehat{\mu}(x)$ is zero, because

$$\mathrm{E}\{\widehat{\mu}(x) - \mu(x)\} = \mathrm{E}\left(B_x\widehat{\theta} - B_x\theta\right\} = 0,$$

so $\widehat{\mu}(x)$ is (unconditionally) unbiased for $\mu(x)$.

☐ To account for the added variability due to the conditional bias, we average the conditional mean square error of $\widehat{\mu}(x)$, i.e.,

$$\mathrm{E}[\{\widehat{\mu}(x) - \mu(x)\}^2 \mid b] = \mathrm{var}\{\widehat{\mu}(x) \mid b\} + \mathrm{E}\{\widehat{\mu}(x) - \mu(x) \mid b\}^2,$$

over the distribution of b, and this is

$$\mathrm{E}_b\left(\mathrm{E}[\{\widehat{\mu}(x) - \mu(x)\}^2 \mid b]\right) = \mathrm{var}\{\widehat{\mu}(x) - \mu(x)\} = B_x \mathrm{var}(\widehat{\theta} - \theta) B_x^{\mathrm{T}},$$

which equals

$$\sigma^2 B_x (B^{\mathrm{T}}B + \lambda D)^{-1} B_x^{\mathrm{T}}.$$

☐ Hence

$$\frac{\widehat{\mu}(x) - \mu(x)}{\sqrt{\mathrm{var}\{\widehat{\mu}(x) - \mu(x)\}}} \sim \mathcal{N}(0, 1),$$

yielding a confidence interval $\widehat{\mu}(x) \pm \sqrt{\widehat{\mathrm{var}}\{\widehat{\mu}(x) - \mu(x)\}} \times t_{\mathrm{df}_{\mathrm{res}}}(1 - \alpha)$ for $\mu(x)$.

Comparison

Lemma 28 *Let $\widehat{\mu} = B\widehat{\theta}$ be the BLUP of $\mu = B\theta$, which contains the values of $\mu(x)$ corresponding to the sample (x_j, y_j) for $j = 1, \ldots, n$. Then setting $B_x = B$ above, we have*

$$\mathrm{var}(\widehat{\mu} - \mu) = \sigma^2 S_\lambda, \quad \mathrm{var}(\widehat{\mu} - \mu \mid b) = \sigma^2 S_\lambda S_\lambda^{\mathrm{T}},$$

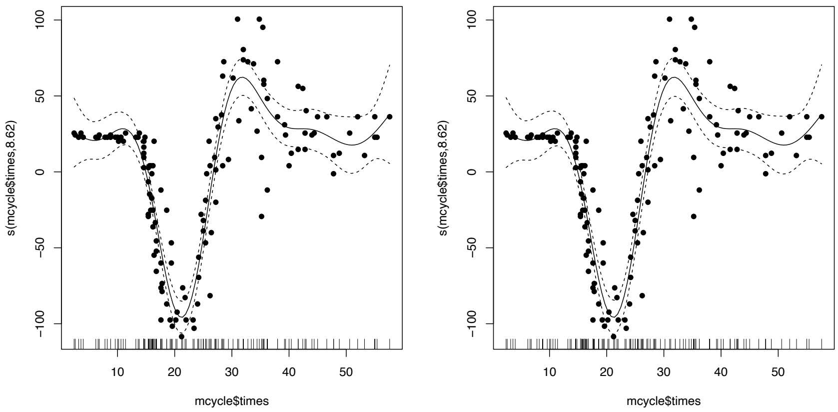
where $S_\lambda = B(B^{\mathrm{T}}B + \lambda D)^{-1}B^{\mathrm{T}}$.

☐ The eigenvalues of S_λ lie in $[0, 1]$, with p of them equal to 1, q of them in the interval $(0, 1)$, and the rest equal to zero, so q of the eigenvalues of $S_\lambda S_\lambda^{\mathrm{T}}$ are smaller than those of S_λ.

☐ Since $S_\lambda S_\lambda^{\mathrm{T}}$ corresponds to 'smoothing twice', the result is less variable than for S_λ, so the unconditional confidence limits will be wider.

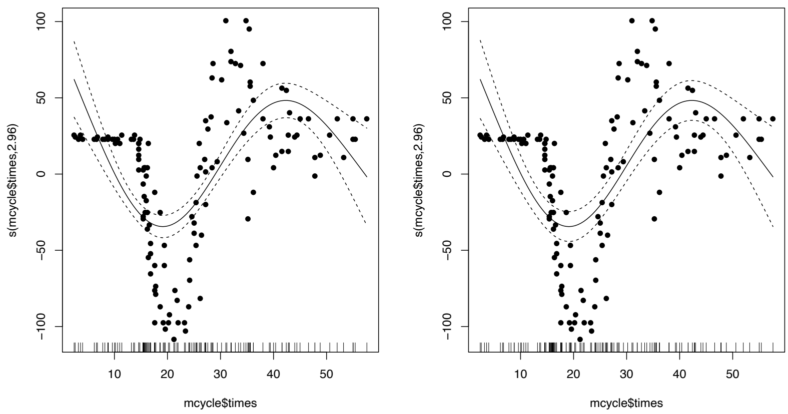
Motorcycle data

Splines fits with conditional (left) and unconditional (right) uncertainty bands. Any differences are almost invisible.



Motorcycle data

Splines fits with conditional (left) and unconditional (right) uncertainty bands, with too few degrees of freedom. Differences between the confidence bands are visible.



Simultaneous confidence bands

- For simultaneous (or overall) bands on $\mathcal{S}$, we seek functions of the data, $L(x)$ and $U(x)$, such that

$$\mathrm{P}\left\{L(x) \le \mu(x) \le U(x), x \in \mathcal{S}\right\} = 1 - 2\alpha.$$

- We approximate $\mathcal{S}$ by a grid $\mathcal{S}_+$ of M points with corresponding $M \times (p+q)$ basis matrix B_+, and BLUP $\widehat{\mu}_+ = B_+\widehat{\theta}$.
- We note that

$$\widehat{\mu}_+ - \mu_+ \approx B_+(\widehat{\theta} - \theta), \quad \widehat{\theta} - \theta \sim \mathcal{N}\left\{0, \sigma^2 B_+(B^{\mathrm{T}}B + \lambda D)^{-1}B_+^{\mathrm{T}}\right\},$$

simulate many times from the fitted model to obtain the $(1-\alpha)$ quantile $m_{1-\alpha}$ of

$$\max_{x\in\mathcal{S}_+}\left|\frac{\widehat{\mu}_+^*(x) - \mu_+^*(x)}{\left\{\widehat{\sigma}^{*2}B_x(B^{\mathrm{T}}B + \widehat{\lambda}^{*2p}D)^{-1}B_x^{\mathrm{T}}\right\}^{1/2}}\right| \approx \sup_{x\in\mathcal{S}}\left|\frac{\widehat{\mu}_+^*(x) - \mu_+^*(x)}{\left\{\widehat{\sigma}^{*2}B_x(B^{\mathrm{T}}B + \widehat{\lambda}^{*2p}D)^{-1}B_x^{\mathrm{T}}\right\}^{1/2}}\right|,$$

where a star indicates something computed from a fit to the simulated data, and then set $L(x)$ and $U(x)$ to be

$$\widehat{\mu}_+ \pm m_{1-\alpha}\left[\mathrm{diag}\left\{\widehat{\sigma}^2 B_+(B^{\mathrm{T}}B + \widehat{\lambda}D)^{-1}B_+^{\mathrm{T}}\right\}\right]^{1/2}, \quad x \in \mathcal{S}_+.$$

Comments

- The computation above uses the marginal distribution of $\widehat{\mu}_+ - \mu_+$, but the bands tend to be a little too wide, so the overall coverage probabilities are too high.
- Using the conditional distribution of $\widehat{\mu}_+ - \mu_+$ fixes this, giving confidence bands with coverage closer to nominal.
- Other approaches are
  - based on the **method of tubes**, which uses a Gaussian process approximation from upcrossing theory, and
  - taking a **Bayesian approach**,

  but they seem not to work as well as using the mixed model formulation of spline smoothing.
- In general it's useful to plot example simulations (picture on board).

Testing parametric fits

- ☐ When writing
$$y \mid b \sim \mathcal{N}_n(X\beta + Zb, \sigma^2 I_n), \quad b \sim \mathcal{N}_q(0, \sigma^2\lambda^{-1} I_q),$$
the matrix X can include variables we don't smooth on, so
  - $X_{n\times(m+p)}$ corresponds to m covariates and p polynomial columns of B,
  - $Z_{n\times q}$ corresponds to penalised columns of B,

  and then testing for no smooth effects corresponds to testing $\lambda = \infty$ or equivalently $\sigma_b^2 = \sigma^2\lambda^{-1} = 0$.
- ☐ The likelihood ratio statistic for this based on the restricted log likelihood ℓ is
$$2\left\{\ell(\widehat{\sigma}^2, \widehat{\sigma}_b^2) - \ell(\widehat{\sigma}_0^2, 0)\right\} \;\dot{\sim}\; \tfrac{1}{2}\chi_0^2 + \tfrac{1}{2}\chi_1^2, \quad n \to \infty,$$
but this approximation may be poor, so using improved asymptotics or simulation from the fitted null model is better.
- ☐ We can also use F-tests, as with local polynomial smoothing, based on comparing the distributions of $y^{\mathrm{T}}(I - H)y$ and $y^{\mathrm{T}}(I - S_\lambda)^{\mathrm{T}}(I - S_\lambda)y$ through an F-statistic.

Comments

- ☐ The mixed model formulation allows an integrated treatment of inference for 'nonparametric' fits, through
  - the least squares estimate $\widehat{\beta} \equiv \widehat{\beta} = (X^{\mathrm{T}}\Delta X)^{-1}X^{\mathrm{T}}\Delta y$ of the parameter β, and
  - the BLUP $\widehat{b} = \Omega_b Z^{\mathrm{T}}\Delta(y - X\widehat{\beta})$ of the random effect b.
- ☐ Both conditional and unconditional analysis of these are possible, with the unconditional analysis giving wider confidence limits that allow for estimation of the bias.
- ☐ Conditional analysis seems to be better for simultaneous (overall) confidence bands.
- ☐ The same framework enables tests of the fit of parametric models, and estimation of derivatives (exercise).

Generalisations

- ☐ We've discussed estimation of a single function $\mu(x)$, but in applications we may have
  - covariates to be treated parametrically,
  - several smooth functions,
  - non-normal response variable,
  - random effects.
- ☐ To include ordinary covariates, we write

$$y \sim (B\theta, \sigma^2 I_n), \quad B\theta = X\beta + Zb,$$

where
  - X represents the ordinary covariates, plus any polynomial columns for smooth components,
  - the 'fixed effects' parameter vector β is not penalized,
  - Z is the basis representation for the smooth function,
  - the 'random effects' vector b is penalized,

and everything 'goes through as before'.

Additive model

- ☐ We suppose

$$y_j \overset{\text{ind}}{\sim} (\mu_1(x_j) + \mu_2(z_j), \sigma^2), \quad j = 1, \ldots, n,$$

where μ_1, μ_2 belong to suitable classes of 'smooth' functions; for example if

$$x \equiv \text{time}, \quad z \equiv \text{space},$$

then μ_1 is defined on $\mathcal{X}_1 \subset \mathbb{R}$ and μ_2 is defined on $\mathcal{X}_2 \subset \mathbb{R}^2$.
- ☐ This is an **additive model**, which might also include ordinary covariates.
- ☐ There is an identifiability problem, since we could map

$$\mu_1(x) \mapsto \mu_1(x) + a, \quad \mu_2(z) \mapsto \mu_2(z) - a, \quad a \in \mathbb{R},$$

and the fitted values would not change, so we must constrain μ_1 and μ_2.
- ☐ As before, we use bases for μ_1 and μ_2, writing

$$y = B\theta + \varepsilon = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} Z_1(x) & Z_2(z) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \varepsilon,$$

where we shall penalise the q_1 elements of b_1 and the q_2 elements of b_2.

Estimation

- We minimise the penalised sum of squares

$$(y - B\theta)^{\mathrm{T}}(y - B\theta) + \theta^{\mathrm{T}} D_\lambda \theta,$$

where

$$D_\lambda = \lambda_1 D_1 + \lambda_2 D_2, \quad D_r = \mathrm{diag}(0, \dots, 0, 1, \dots, 1, 0, \dots, 0), \quad r = 1, 2,$$

and the 1s on the diagonals of D_r correspond to the coefficients to be penalised.

- The solution for fixed λ_1, λ_2 is the same as before, i.e.,

$$\widehat{\theta} = (B^{\mathrm{T}}B + D_\lambda)^{-1} B^{\mathrm{T}} y,$$

where $\widehat{\theta} \equiv \widehat{\theta}_\lambda$, and we've now absorbed the smoothing parameters into D_λ, giving fitted values in terms of the smoothing matrix,

$$\widehat{y} = B\widehat{\theta} = B(B^{\mathrm{T}}B + D_\lambda)^{-1} B^{\mathrm{T}} y = S_\lambda y.$$

- Now we must decide
  - how many degrees of freedom for each smooth?
  - how to select the smoothing parameters?

Effective degrees of freedom

- Denote the usual (unpenalized) least squares estimate of $\theta_{(p+q)\times 1}$ by

$$\widehat{\theta} = (B^{\mathrm{T}}B)^{-1} B^{\mathrm{T}} y,$$

and note that we can write

$$\widehat{\theta}_\lambda = S_\lambda y = (B^{\mathrm{T}}B + D_\lambda)^{-1} B^{\mathrm{T}} y = (B^{\mathrm{T}}B + D_\lambda)^{-1} B^{\mathrm{T}} B \widehat{\theta} = P_\lambda \widehat{\theta},$$

say, where P_λ shows how penalisation shrinks $\widehat{\theta}$ towards $\widehat{\theta}_\infty = (\widehat{\beta}^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$.

- If $\lambda \approx 0$, then $P_\lambda \approx I_{p+q}$ and the degrees of freedom of the two fits are both $\approx p + q$, but as $\lambda \to \infty$, P_λ tends to the projection matrix onto the column space of $X_{n\times p}$.
- On slide 175 with just one smooth term we defined

$$\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda) = \mathrm{tr}(P_\lambda) = \sum_{r=1}^{p+q} P_{\lambda,rr} \in (p, p+q),$$

which gives the usual definition for a linear model.

- With several smooth terms, $\theta^{\mathrm{T}} = (\theta_1^{\mathrm{T}}, \dots, \theta_M^{\mathrm{T}})$, say, we define the **effective degrees of freedom** df_{λ_m} associated to the mth smooth as being the sum of those $P_{\lambda,rr}$ that correspond to the elements of θ_m in θ.

Selection of λ

- ☐ Line up the usual suspects:
  - – cross-validation sum of squares, $\mathrm{CV}(\lambda)$,
  - – generalized cross-validation sum of squares, $\mathrm{GCV}(\lambda)$, and
  - – REML, $\ell(\sigma^2, \lambda)$.
- ☐ The first two involve a grid search over values of (log-)λ, and can be expensive.
- ☐ We discuss implementation of the third, which in principle could be cheaper, later.
- ☐ Our previous ideas on uncertainty estimation for the parametric and smooth terms go through ...

Identifiability

- ☐ The identifiability problem is solved by **centering** the fitted smooth, i.e., enforcing

$$1_n^{\mathrm{T}} Z_{n\times q} b_{q\times 1} = 0$$

for each smooth term.

- ☐ In general we can use a QR decomposition. If $C_{m\times q} b_{q\times 1} = 0_{m\times 1}$, with $m < q$, write

$$C_{q\times m}^{\mathrm{T}} = Q_{q\times q} R_{q\times m} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix},$$

where Q is orthogonal,
  - – Q_1 has dimension $q \times m$,
  - – Q_2 has dimension $q \times (q-m)$, and
  - – R_1 has dimension $m \times m$ and is upper triangular.

Then if we set $b_{q\times 1} = Q_2 b'_{(q-m)\times 1}$, we have

$$Cb = R^{\mathrm{T}} Q^{\mathrm{T}} b = \begin{pmatrix} R_1^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} Q_1^{\mathrm{T}} \\ Q_2^{\mathrm{T}} \end{pmatrix} Q_2 b' = \begin{pmatrix} R_1^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ I_{q-m} \end{pmatrix} b' = 0.$$

- ☐ Thus the constraint is satisfied if we replace $Z_{n\times q}$ by $(ZQ_2)_{n\times(q-1)}$; this reduces b to dimension $(q-1)\times 1$. This is efficiently implemented using Householder operations.

Numerical aspects

☐ We need to minimise

$$(y - B\theta)^{\mathrm{T}}(y - B\theta) + \theta^{\mathrm{T}} D_\lambda \theta,$$

where n and q may be large. Let A_λ be a square root matrix of D_λ, i.e., $A_\lambda^{\mathrm{T}} A_\lambda = D_\lambda$, and note that

$$\left\| \begin{pmatrix} y \\ 0 \end{pmatrix}_{(n+q)\times 1} - \begin{pmatrix} B \\ A_\lambda \end{pmatrix}_{(n+q)\times q} \theta_{q\times 1} \right\|^2 = \|y - B\theta\|^2 + \theta^{\mathrm{T}} D_\lambda \theta$$

has the form of an augmented least squares probem.

☐ A_λ can be found using Choleski decomposition with $\mathcal{O}(q^3)$ operations.

☐ The least squares problem can be solved stably and efficiently using standard orthogonal matrix methods, even for very large $n + q$.

☐ To estimate λ by generalized cross-validation: set $k = 1$, $\lambda_1 = 10^{-8}$, $k_{\max} = 60$, then while $k < k_{\max}$,

- compute $\mathrm{GCV}(\lambda_k)$ using the methods above,
- set $\lambda_k \leftarrow 1.5\lambda_k$,
- set $k \leftarrow k + 1$,

☐ stop and output fit for best value of λ

Numerical aspects: REML

☐ With the linear model

$$y \mid b \sim \mathcal{N}_n(X\beta + Zb, \sigma^2 I_n), \quad b \sim \mathcal{N}_k(0, \sigma^2 D_\lambda),$$

we get restricted log likelihood

$$\ell(\sigma^2, \lambda) \equiv -\tfrac{1}{2}\log|U_\lambda| - \tfrac{1}{2}\log|X^{\mathrm{T}} U_\lambda^{-1} X| - \frac{1}{2\sigma^2} y^{\mathrm{T}} P_\lambda y - \frac{n-q}{2}\log\sigma^2,$$

where $U_\lambda = Z D_\lambda Z^{\mathrm{T}} + I_n$ and $P_\lambda = U_\lambda^{-1} - U_\lambda^{-1} X (X^{\mathrm{T}} U_\lambda^{-1} X)^{-1} X^{\mathrm{T}} U_\lambda^{-1}$.

☐ Given λ, there is an explicit formula for $\widehat{\sigma}^2$, so we finally seek

$$\widehat{\lambda} = \mathrm{argmax}\left\{ -\tfrac{1}{2}\log|U_\lambda| - \tfrac{1}{2}\log|X^{\mathrm{T}} U_\lambda^{-1} X| - \frac{n-q}{2}\log y^{\mathrm{T}} P_\lambda y \right\},$$

which involves some unpleasant matrix differentials, e.g.,

$$\frac{\partial Q^{-1}}{\partial\lambda} = -Q^{-1}\frac{\partial Q}{\partial\lambda}Q^{-1}, \quad \frac{\partial \log|Q|}{\partial\lambda} = \mathrm{tr}\left(Q^{-1}\frac{\partial Q}{\partial\lambda}\right)$$

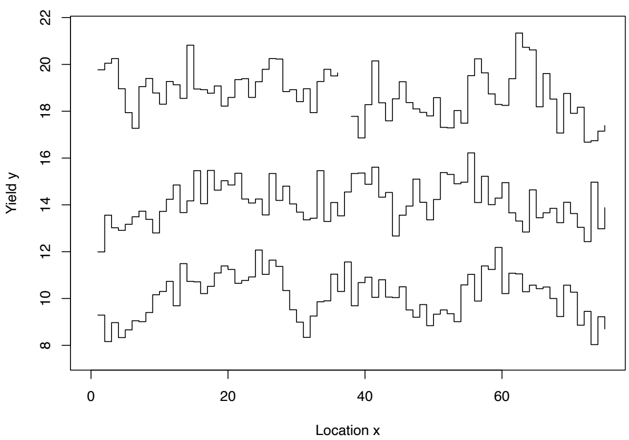
but is just a function of λ. Reasonably efficient computational methods are available ...

Example: Spring barley data

Plot yield at harvest for 75 varieties of spring barley sown in 3 blocks each of 75 plots:

Location t	Block 1		Block 2		Block 3	
	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example: Spring barley data



Yield as a function of location for the three blocks, with yields for blocks 2 and 3 offset by the addition of 4 and of 8 respectively. Value 37 in block 3 (corresponding to variety 27) is missing.

Example: Spring barley data

- We fit a model with parametric variety effects and smooth effects for the fertility patterns in the blocks,
$$y_{n\times 1} \sim (X_{n\times 75}\beta_{75\times 1} + Z_1\theta_1 + Z_2\theta_2 + Z_3\theta_3, \sigma^2 I_n),$$
where
  - $n = 224$, as one of the responses is missing,
  - X is a matrix of indicators (0/1) of which variety is in which plot in each block,
  - β are the variety effects, with the model parametrized without an overall mean,
  - Z_m of dimension $n \times (p_m + q_m)$ corresponds to the basis functions for the smooth in block m, and
  - θ_m are of dimensions $(p_m + q_m) \times 1$, for $m = 1, 2, 3$, corresponding to the smooth effects, and
  - $p_m + q_m = 9$ by default (after centering) when using `gam` in `R` package `mgcv`.
- Taking $p_m = 2$ would correspond to null smooth $\beta_0 + \beta_1 x$ for each block (i.e., linear fertility pattern), but the identifiability constraints impose $\beta_0 = 0$. Hence in fact $p_m = 1$ for a linear baseline smooth and the degrees of freedom for the smooth terms lie in $[1, 9]$ (see slide 206).

Example: Spring barley data

```
library(SMPracticals)
data(barley)

library(mgcv)

# ML fit of variety as fixed effect, with GCV estimation of lambdas,
# with splines for fertility gradients within each block

fit.gcv <- gam(y~Variety-1+s(Location,by=Block),data=barley)

# fit of variety as fixed effect, with REML estimation of lambdas,
# with splines for fertility gradients within each block

fit <- gam(y~Variety-1+s(Location,by=Block),method="REML",data=barley)

# REML fit with variety as a random effect and splines for fertilities

fit.re <- gam(y~s(Variety,bs="re")+s(Location,by=Block),method="REML",
                    data=barley)
```

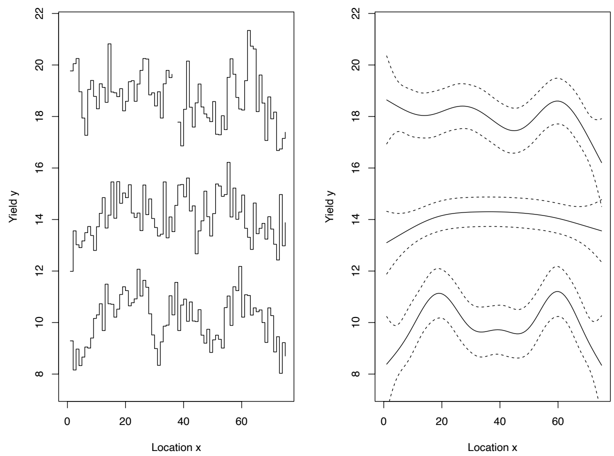
Example: Spring barley data

- ☐ Using GCV the smooths have $\mathrm{df}_\lambda = 8.3$, 6.8, 6.3, with $\widehat{\sigma} = 0.65$ and $\mathrm{AIC} = 513.1$, the residual degrees of freedom is $224 - 75 - 8.3 - 6.8 - 6.3 \approx 130.6$, with SEs around 0.4 for the estimated variety effects (0.54 for variety 27).
- ☐ Using REML the smooths have $\mathrm{df}_\lambda = 7.2$, 3, 6.1, with $\widehat{\sigma} = 0.66$ and $\mathrm{AIC} = 518.3$, the residual degrees of freedom is 132.7, with SEs around 0.4 for the estimated variety effects (0.53 for variety 27).
- ☐ The estimated smoothing parameters are $\widehat{\lambda}_1 = 0.0029$, $\widehat{\lambda}_2 = 0.18$ and $\widehat{\lambda}_3 = 0.0078$.
- ☐ The effective degrees of freedom for the smooth terms, with the totals:

Block					$P_{\lambda,rr}$					Total
1	1.00	1.07	0.90	0.7	0.65	0.17	0.38	1.31	1	7.18
2	0.61	0.21	0.12	-0.2	0.03	-0.26	0.01	1.49	1	3.00
3	0.99	1.04	0.76	0.4	0.41	-0.18	0.18	1.47	1	6.07

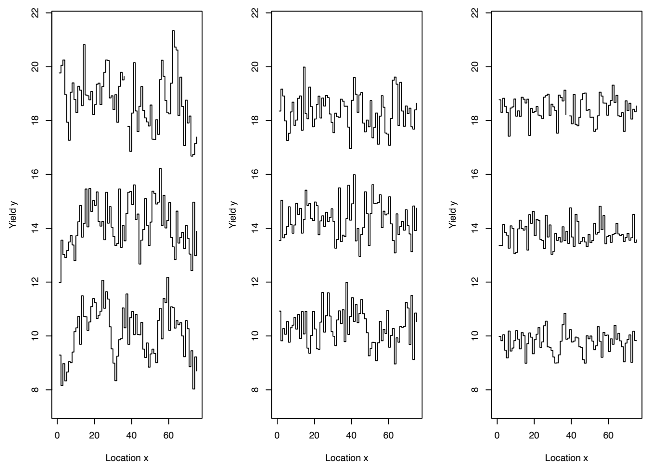
- ☐ The $P_{\lambda,rr}$ need not be positive, though their total for each smooth is positive.
- ☐ In applications it would be wise to check whether increasing q_m would lead to very different fits.

Example: Spring barley data



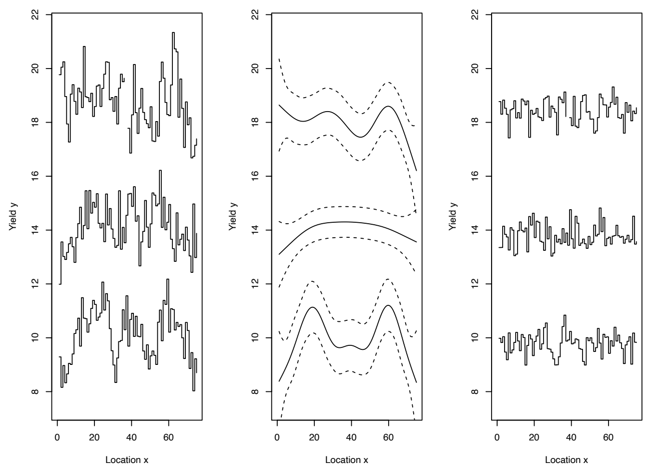
Left: data (offset by adding 4 and 8 to blocks 2 and 3).
Right: estimated fertility patterns (with estimated df 7.2, 3, 6.1) and 95% unconditional pointwise confidence intervals, fitted using REML. The intervals are wider for blocks 1 and 3.

Example: Spring barley data



Left: data (offset by adding 4 and 8 to blocks 2 and 3).
Center: Estimated variety effects (also offset)
Right: residuals (also offset, and showing serial autocorrelation?)

Example: Spring barley data



Left: data (offset by adding 4 and 8 to blocks 2 and 3).
Center: estimated fertility patterns (REML), also offset.
Right: residuals.

Example: Spring barley data

- ☐ Should the varieties be treated as randomly selected from a population of varieties?
- ☐ If so, we use the same basis matrix X as in the previous model, but add a penalty matrix $\lambda_\beta D_\beta$ and minimise the penalised sum of squares

$$(y - B\theta)^{\mathrm{T}}(y - B\theta) + \theta^{\mathrm{T}} D_\lambda \theta,$$

where

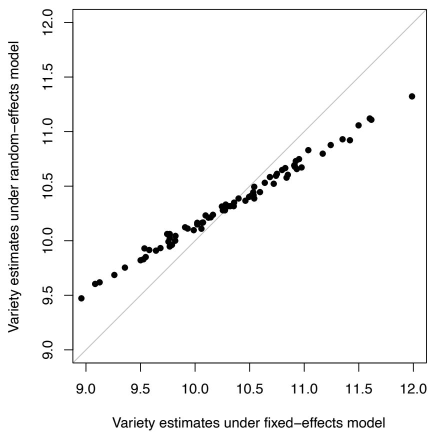
$$D_\lambda = \lambda_\beta D_\beta + \lambda_1 D_1 + \lambda_2 D_2 + \lambda_3 D_3,$$

where $D_\beta = \mathrm{diag}(I_{75}, 0)$.

- ☐ The effective degrees of freedom for this model are 44.8 for β and 7.5, 3.9 and 6.4 for the splines.
- ☐ The optimal smoothing parameters are $\widehat{\lambda}_\beta = 1.76$, $\widehat{\lambda}_1 = 0.0027$, $\widehat{\lambda}_2 = 0.073$ and $\widehat{\lambda}_3 = 0.0070$.
- ☐ The fixed-effects model has 75 degrees of freedom for β, so this is substantial shrinkage; the estimated standard deviation drops from 0.65 to 0.39.
- ☐ The estimates under the random-effects model have standard errors around 0.31 (0.36 for variety 27), compared to 0.41 (0.54 for variety 27) for the fixed-effects model.
- ☐ The next slide compares the estimates.

Example: Spring barley data

Comparison of estimated variety effects under fixed-effects and random-effects models:



Discussion

- ☐ The basic ideas of spline smoothing extend to the additive model:
  - use of basis functions to represent smooth curves;
  - penalization of parameters/random effects to impose smoothness;
  - REML/GCV estimation of smoothing parameters λ;
  - linear estimation using shrinkage matrix S_λ;
  - effective degrees of freedom defined using eigenvalues of S_λ,

  with
  - inclusion of (unpenalized) parameters for covariate effects;
  - different smoothers for different variables, penalised individually.
- ☐ Penalization corresponds to Bayesian model, with smoothing parameters estimated from the data (empirical Bayes estimation).

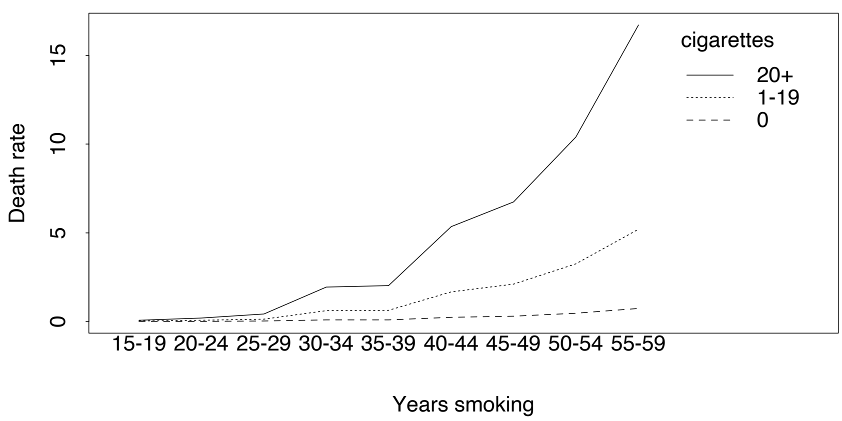
Smoking data

Table 2: Lung cancer deaths in British male physicians (Doll and Hill, 1952). The table gives man-years at risk T/number of cases y of lung cancer, cross-classified by years of smoking t, taken to be age minus 20 years, and number of cigarettes smoked per day, d.

Years of smoking t	Daily cigarette consumption d						
	Nonsmokers	1–9	10–14	15–19	20–24	25–34	35+
15–19	10366/1	3121	3577	4317	5683	3042	670
20–24	8162	2937	3286/1	4214	6385/1	4050/1	1166
25–29	5969	2288	2546/1	3185	5483/1	4290/4	1482
30–34	4496	2015	2219/2	2560/4	4687/6	4268/9	1580/4
35–39	3512	1648/1	1826	1893	3646/5	3529/9	1336/6
40–44	2201	1310/2	1386/1	1334/2	2411/12	2424/11	924/10
45–49	1421	927	988/2	849/2	1567/9	1409/10	556/7
50–54	1121	710/3	684/4	470/2	857/7	663/5	255/4
55–59	826/2	606	449/3	280/5	416/7	284/3	104/1

Smoking data

Lung cancer deaths in British male physicians. The figure shows the rate of deaths per 1000 man-years at risk, for each of three levels of daily cigarette consumption.



Smoking data

☐ Suppose number of deaths y has Poisson distribution, mean $T\lambda(d,t)$, where T is man-years at risk, d is number of cigarettes smoked daily and t is time smoking (years).

☐ Take

$$\lambda(d,t) = \beta_0 t^{\beta_1}\left(1 + \beta_2 d^{\beta_3}\right):$$

– background rate of lung cancer is $\beta_0 t^{\beta_1}$ for non-smoker,

– additional risk due to smoking d cigarettes/day is $\beta_2 d^{\beta_3}$.

☐ With $x_j = (T_j, d_j, t_j)$, can write this as

$$\begin{aligned} y_j &\sim \text{Poiss}\{\mu(\beta; x_j)\}, \\ \mu(\beta; x) &= T\beta_0 t^{\beta_1}\left(1 + \beta_2 d^{\beta_3}\right), \quad j = 1, \ldots, n: \end{aligned}$$

a nonlinear model with Poisson-distributed response.

Challenger data

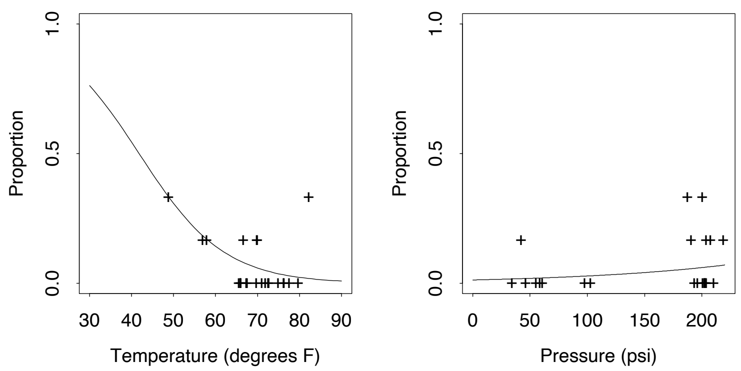
Challenger data

Table 3: O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch)

Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
⋮				
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200
61-I	28/1/86	—	31	200

Challenger data

Figure 1: O-ring thermal distress data. The left panel shows the proportion of incidents as a function of joint temperature, and the right panel shows the corresponding plot against pressure. The x-values have been jittered to avoid overplotting multiple points. The solid lines show the fitted proportions of failures under a logistic regression model.



Comments

- ☐ Linear model $y \sim (X\beta, \sigma^2 I_n)$
  - applicable for continuous response $y \in \mathbb{R}$
  - assumes linear dependence of mean response $\mathrm{E}(y)$ on covariates X
  - sometimes assumes y normal
- ☐ Lots of data not like this
- ☐ Need extensions for
  - nonlinear dependence on covariates
  - arbitrary response distribution (binomial, Poisson, exponential, ...)
  - dependent responses
  - variance non-constant (and related to mean?)
  - censoring, truncation, ...
  - ...

Simple fixes

- ☐ Just fit a linear model anyway
  - Might work as an approximation, but usually extrapolates really badly.
- ☐ Fit a linear model to transformed responses
  - E.g., take variance-stabilising transformation for y, such as $2\sqrt{y}$ when y is Poisson
  - Can be helpful, but usually the obvious transformation can't give linearity.
- ☐ Instead we attempt to fit the model using likelihood estimation.

Likelihood

Definition 29 *Let y be a data set, assumed to be the realisation of a random variable $Y \sim f(y;\theta)$, where the unknown parameter θ lies in the parameter space $\Omega_\theta \subset \mathbb{R}^p$. Then the* **likelihood** *(for θ based on y) and the corresponding* **log likelihood** *are*

$$L(\theta) = L(\theta; y) = f_Y(y;\theta), \quad \ell(\theta) = \log L(\theta), \quad \theta \in \Omega_\theta.$$

The **maximum likelihood estimate** *(MLE) $\widehat{\theta}$ satisfies $\ell(\widehat{\theta}) \geq \ell(\theta)$, for all $\theta \in \Omega_\theta$.*
Often $\widehat{\theta}$ is unique and in many cases it satisfies the **score (or likelihood) equation**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

which is interpreted as a vector equation of dimension $p \times 1$ if θ is a $p \times 1$ vector.
The **observed information** *and* **expected (Fisher) information** *are defined as*

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}, \quad I(\theta) = \mathrm{E}\{J(\theta)\};$$

these are $p \times p$ matrices if θ has dimension p.

Maximum likelihood estimator

- In large samples from a **regular model** in which the true parameter is $\theta^0_{p\times 1}$, the maximum likelihood estimator $\widehat{\theta}$ has an approximate normal distribution,

  $$\widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_p\left\{\theta^0, J(\widehat{\theta})^{-1}\right\},$$

  so we can compute an approximate $(1-2\alpha)$ confidence interval for the rth parameter θ^0_r as

  $$\widehat{\theta}_r \pm z_\alpha v_{rr}^{1/2},$$

  where v_{rr} is the rth diagonal element of the matrix $J(\widehat{\theta})^{-1}$.
- This is easily implemented:
  - we code the negative log likelihood $-\ell(\theta)$ (and check the code carefully!);
  - we minimise $-\ell(\theta)$ numerically, ensuring that the minimisation routine returns $\widehat{\theta}$ and the Hessian matrix $J(\widehat{\theta}) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}|_{\theta=\widehat{\theta}}$
  - we compute $J(\widehat{\theta})^{-1}$, and use the square roots of its diagonal elements, $v_{11}^{1/2}, \ldots, v_{dd}^{1/2}$, as standard errors for the corresponding elements of $\widehat{\theta}$.

Aside: Regular model

We say that a statistical model $f(y;\theta)$ is **regular (for likelihood inference)** if

1. the true value θ^0 of θ is interior to the parameter space $\Omega_\theta \subset \mathbb{R}^p$;
2. the densities defined by any two different values of θ are distinct;
3. there is an open set $\mathcal{I} \subset \Omega_\theta$ containing θ^0 within which the first three derivatives of the log likelihood with respect to elements of θ exist almost surely, and

$$|\partial^3 \log f(Y_j;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| \le g(Y_j)$$

   uniformly for $\theta \in \mathcal{I}$, where $0 < \mathrm{E}_0\{g(Y_j)\} = K < \infty$; and
4. for $\theta \in \mathcal{I}$ we can interchange differentation with respect to θ and integration, that is,

$$\frac{\partial}{\partial\theta}\int f(y;\theta)\, dy = \int \frac{\partial f(y;\theta)}{\partial\theta}\, dy, \quad \frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\int f(y;\theta)\, dy = \int \frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\, dy.$$

The results are also true under weaker conditions, for non-identically distributed and dependent data.

Aside: Comments on regular models

Condition

1. is needed so that $\widehat{\theta}$ can lie 'on both sides' of θ^0 and hence can have a limiting normal distribution, once standardized—**fails**, for example, if θ has a discrete component (e.g. changepoint $\gamma \in \{1,\ldots,n\}$);
2. is needed to be able to identify the model on the basis of the data;
3. ensures the validity of Taylor series expansions of $\ell(\theta)$—not usually a problem;
4. ensures that the score statistic has a limiting normal distribution—can **fail** in some models — sometimes good news, leading to faster convergence than $n^{-1/2}$.

All the above assumes the postulated model is correct! — there is a literature on what happens when we fit the wrong model, or if the parameter dimension increases with n, or ... usually there are no generic results for such cases.

Likelihood ratio statistic

- Model $f_B(y)$ is **nested** within model $f_A(y)$ if A reduces to B on restricting some parameters:
  - for example, the model $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is nested within the model $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, because the first is obtained from the second by setting $\mu = 0$;
  - the maximised log likelihoods satisfy $\widehat{\ell}_A \geq \widehat{\ell}_B$, because the more comprehensive model A contains the simpler model B.
- The **likelihood ratio statistic** for comparing them is

$$W = 2(\widehat{\ell}_A - \widehat{\ell}_B).$$

- If the model is regular, the simpler model is true, and A has q more parameters than B, then

$$W \dot{\sim} \chi^2_q.$$

- This implicitly assumes that ML inference for model A is OK, so that the approximation $\widehat{\theta}_A \dot{\sim} \mathcal{N}\{\theta_A, J_A(\widehat{\theta}_A)^{-1}\}$ is adequate.

Profile log likelihood

- Consider a regular log likelihood $\ell(\psi, \lambda)$, where the **parameter of interest** ψ is variation independent of the **nuisance parameter** λ, i.e., $(\psi, \lambda) \in \Omega_\psi \times \Omega_\lambda$, and the overall MLE is $(\widehat{\psi}, \widehat{\lambda})$.
- For a confidence set for ψ, without reference to λ, we use the **profile log likelihood**

$$\ell_{\text{p}}(\psi) = \max_{\lambda \in \Omega_\lambda} \ell(\psi, \lambda) = \ell(\psi, \widehat{\lambda}_\psi),$$

  say, and, based on the limiting distribution of the likelihood ratio statistic, take as $(1 - 2\alpha)$ confidence region the set

$$\left\{\psi \in \Omega_\psi : 2\{\ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi, \widehat{\lambda}_\psi)\} \leq \chi^2_{\dim \psi}(1 - 2\alpha)\right\}.$$

- When ψ is scalar, this yields

$$\left\{\psi \in \Omega_\psi : \ell(\psi, \widehat{\lambda}_\psi)\} \geq \ell(\widehat{\psi}, \widehat{\lambda}) - \tfrac{1}{2}\chi^2_1(1 - 2\alpha)\right\},$$

  and $\frac{1}{2}\chi^2_1(0.95) = 1.92$.
- Such intervals are generally better than the standard interval $\widehat{\psi} \pm z_\alpha \text{SE}$, particularly when the distribution of $\widehat{\psi}$ is asymmetric, but require more computation, since they involve many maximisations of ℓ.

Model setup

- Independent random variables $Y_1, \ldots, Y_n$, with observed values $y_1, \ldots, y_n$, and covariates $x_1, \ldots, x_n$.
- Suppose that probability density of Y_j is $f(y_j; \eta_j, \phi)$, where $\eta_j = \eta(\beta, x_j)$, and ϕ is common to all models.
- Log likelihood is

$$\ell(\beta, \phi) = \sum_{j=1}^{n} \ell_j(\beta, \phi) = \sum_{j=1}^{n} \log f\{y_j; \eta(\beta, x_j), \phi\}.$$

- More generally, just let $\ell_j(\beta, \phi)$ denote the log likelihood contribution from the $j\text{th}$ observation.
- Suppose ϕ known (for now), suppress it, and estimate β.

Example 30 (Normal regression model) *Express the normal regression model in the terms above.*

Note to Example 30

Here $Y_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$ with $\mu_j = \eta_j = \eta(x_j; \beta)$, so obviously

$$\eta_j = \eta(x_j; \beta), \quad \phi = \sigma^2, \quad \ell_j \equiv -\tfrac{1}{2}\{(y_j - \eta_j)^2/\phi + \log \phi\}.$$

Iterative weighted least squares (IWLS)

- General approach for estimation in regression models, based on Newton–Raphson iteration
- Assume that ϕ is fixed, and write

$$\ell(\beta) = \sum_{j=1}^{n} \ell_j\{\eta_j(\beta)\}.$$

- MLEs $\widehat{\beta}$ usually satisfy

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta_r} = 0, \qquad r = 1, \ldots, p,$$

  or equivalently

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} \frac{\partial \ell}{\partial \eta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} u(\widehat{\beta}) = 0, \tag{13}$$

  where $u(\beta)$ is $n \times 1$ vector with $j\text{th}$ element $\partial \ell / \partial \eta_j$.

IWLS II

- Newton–Raphson update step:
$$\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wz,$$
where
$$\begin{aligned} X_{n\times p} &= \partial\eta/\partial\beta^{\mathrm{T}}, \quad \text{(design matrix)} \\ W_{n\times n} &= \mathrm{diag}\{\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)\}, \quad \text{(weights)} \\ z_{n\times 1} &= X\beta + W^{-1}u, \quad \text{(adjusted dependent variable)} \end{aligned}$$
- Thus to obtain MLEs $\widehat{\beta}$ we use the **IWLS algorithm**:
- take an initial $\widehat{\beta}$. Repeat
  - compute X, W, u, z;
  - compute new $\widehat{\beta}$ and replace the preceding value;

  until changes in $\ell(\widehat{\beta})$ (or, sometimes, $\widehat{\beta}$, or both) are lower than some tolerance.
- Sometimes a line search is added, if $\ell(\widehat{\beta}_{\mathrm{new}}) < \ell(\widehat{\beta}_{\mathrm{old}})$: i.e., we half the step length and try again.

Derivation of IWLS algorithm

- To find the maximum likelihood estimate $\widehat{\beta}$ starting from a trial value β, we make a Taylor series expansion in (13), to obtain

$$\frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta) + \left\{ \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \frac{\partial^2 \ell_j(\beta)}{\partial \eta_j^2} \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}} + \sum_{j=1}^{n} \frac{\partial^2 \eta_j(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} u_j(\beta) \right\} (\widehat{\beta} - \beta) \doteq 0. \qquad (14)$$

  If we denote the $p \times p$ matrix in braces on the left by the $p \times p$ matrix $-J(\beta)$, assumed invertible, we can rearrange (14) to obtain

$$\widehat{\beta} \doteq \beta + J(\beta)^{-1} \frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta). \qquad (15)$$

  This suggests that maximum likelihood estimates may be obtained by starting from a particular β, using (15) to obtain $\widehat{\beta}$, then setting β equal to $\widehat{\beta}$, and iterating (15) until convergence. This is the Newton–Raphson algorithm applied to our particular setting. In practice it can be more convenient to replace $J(\beta)$ by its expected value

$$I(\beta) = \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \mathrm{E}\left(-\frac{\partial^2 \ell_j}{\partial \eta_j^2} \right) \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}};$$

  the other term vanishes because $\mathrm{E}\{u_j(\beta)\} = 0$. We write

$$I(\beta) = X(\beta)^{\mathrm{T}} W(\beta) X(\beta), \qquad (16)$$

  where $X(\beta)$ is the $n \times p$ matrix $\partial \eta(\beta)/\partial \beta^{\mathrm{T}}$ and $W(\beta)$ is the $n \times n$ diagonal matrix whose jth diagonal element is $\mathrm{E}(-\partial^2 \ell_j / \partial \eta_j^2)$.

- If we replace $J(\beta)$ by $X(\beta)^{\mathrm{T}} W(\beta) X(\beta)$ and reorganize (15), we obtain

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W (X\beta + W^{-1} u) = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z, \qquad (17)$$

  say, where the dependence of the terms on the right on β has been suppressed. That is, starting from β, the updated estimate $\widehat{\beta}$ is obtained by weighted linear regression of the $n \times 1$ vector **adjusted dependent variable**

$$z = X(\beta)\beta + W(\beta)^{-1} u(\beta)$$

  on the columns of $X(\beta)$, using weight matrix $W(\beta)$. The maximum likelihood estimates are obtained by repeating this step until the log likelihood, the estimates, or more often both are essentially unchanged. The variable z plays the role of the response or dependent variable in the weighted least squares step.

- Often the structure of a model simplifies the estimation of an unknown value of ϕ. It may be estimated by a separate step between iterations of $\widehat{\beta}$, by including it in the step (15), or from the profile log likelihood $\ell_{\mathrm{p}}(\phi)$.

Examples

Example 31 (Normal nonlinear model) *Give the components of the IWLS algorithm for the normal nonlinear model.*

Note to Example 31

- Here the mean of the jth observation is $\eta_j = \eta(x_j; \beta)$. The log likelihood contribution $\ell_j(\eta_j)$ is

$$\ell_j(\eta_j, \sigma^2) \equiv -\tfrac{1}{2}\left\{\log\sigma^2 + \frac{1}{\sigma^2}(y_j - \eta_j)^2\right\},$$

so

$$u_j(\eta_j) = \frac{\partial \ell_j}{\partial \eta_j} = \frac{1}{\sigma^2}(y_j - \eta_j), \qquad \frac{\partial^2 \ell_j}{\partial \eta_j^2} = -\frac{1}{\sigma^2};$$

the jth element on the diagonal of W is the constant σ^{-2}.
The jth row of the matrix $X = \partial\eta/\partial\beta^{\mathrm{T}}$ is $(\partial\eta_j/\partial\beta_0, \ldots, \partial\eta_j/\partial\beta_{p-1})$, and as η_j is nonlinear as a function of β, X depends on β.
After some simplification, we see that the new value for $\widehat{\beta}$ given by (17) is

$$\widehat{\beta} \doteq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + y - \eta), \qquad (18)$$

where X and η are evaluated at the current β. Here $\eta \neq X\beta$ and (18) must be iterated.
- The log likelihood is a function of β only through the sum of squares, $\mathrm{SS}(\beta) = \sum_{j=1}^n \{y_j - \eta_j(\beta)\}^2$. The profile log likelihood for σ^2 is

$$\ell_{\mathrm{p}}(\sigma^2) = \max_\beta \ell(\beta, \sigma^2) \equiv -\tfrac{1}{2}\left\{n\log\sigma^2 + \mathrm{SS}(\widehat{\beta})/\sigma^2\right\},$$

so the maximum likelihood estimator of σ^2 is $\widehat{\sigma}^2 = \mathrm{SS}(\widehat{\beta})/n$. Although $S^2 = \mathrm{SS}(\widehat{\beta})/(n-p)$ is not unbiased when the model is nonlinear, it turns out to have smaller bias than $\widehat{\sigma}^2$, and is preferable in applications.
- In some cases the error variance depends on covariates, and we write the variance of the jth response as $\sigma_j^2 = \sigma^2(x_j, \gamma)$. Such models may be fitted by alternating iterative weighted least squares updates for β treating γ as fixed at a current value with those for γ with β fixed, convergence being attained when neither estimates nor log likelihood change materially.

Deviance

- Let $\widehat{\eta}_j = \eta_j(\widehat{\beta}, x_j)$, where $\widehat{\beta}$ is MLE of β, giving maximised log likelihood $\ell(\widehat{\beta})$ and $\widehat{\eta}^{\mathrm{T}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_n)$.
- Let $\tilde{\eta}_j$ be the value of η_j that maximises $\log f(y_j; \eta_j)$, and let $\tilde{\eta}^{\mathrm{T}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_n)$. This corresponds to the **saturated model**, with

$$\#\text{parameters in } \eta = \#\text{observations in } y,$$

which will give the largest likelihood possible.
- Define the **scaled deviance**:

$$D = 2\sum_{j=1}^n \{\log f(y_j; \tilde{\eta}_j) - \log f(y_j; \widehat{\eta}_j)\} \geq 0.$$

- Small D implies $\widehat{\eta} \approx \tilde{\eta}$, so model fits well.
- Large D implies poor fit — like $SS(\widehat{\beta})$ in linear model.

Differences of deviances

- ☐ Consider two models:
  - Model A: $\beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ vary freely — MLEs $\widehat{\eta}^A = \eta(\widehat{\beta}^A)$;
  - Model B: $(\beta_1, \ldots, \beta_q) \in \mathbb{R}^q$ vary freely, but $\beta_{q+1}, \ldots, \beta_p$ are fixed — hence q free parameters, MLEs $\widehat{\eta}^B = \eta(\widehat{\beta}^B)$.
- ☐ Model B is **nested within** model A: B can be obtained by restricting A.
- ☐ Likelihood ratio statistic for comparing the models is

$$2(\widehat{\ell}_A - \widehat{\ell}_B) = 2\sum_{j=1}^n \left\{\log f(y_j;\widehat{\eta}_j^A) - \log f(y_j;\widehat{\eta}_j^B)\right\} = D_B - D_A,$$

  and this $\dot{\sim}\ \chi^2_{p-q}$ if the models are regular.
- ☐ If ϕ unknown, replace it by an estimate: same distributional approximations will apply.

Example 32 (Normal linear model) *Find the difference of deviances in the normal linear model.*

Note to Example 32

- ☐ Suppose that the y_j are normal with means η_j and known variance ϕ. Then

$$\log f(y_j;\eta_j,\phi) = -\tfrac{1}{2}\left\{\log(2\pi\phi) + (y_j - \eta_j)^2/\phi\right\}$$

  is maximized with respect to η_j when $\tilde{\eta}_j = y_j$, giving $\log f(y_j;\tilde{\eta}_j,\phi) = -\frac{1}{2}\log(2\pi\phi)$. Therefore the scaled deviance for a model with fitted means $\widehat{\eta}_j$ is

$$D = \phi^{-1}\sum_{j=1}^n (y_j - \widehat{\eta}_j)^2,$$

  which is just the residual sum of squares for the model, divided by ϕ. If $\eta_j = x_j^{\mathrm{T}}\beta$ is the correct normal linear model, the distribution of the residual sum of squares is $\phi\chi^2_{n-p}$, so values of D extreme relative to the χ^2_{n-p} distribution call the model into question.
- ☐ The difference between deviances for nested models A and B in which β has dimensions p and $q < p$,

$$D_B - D_A = \phi^{-1}\sum_{j=1}^n \left\{(y_j - \widehat{\eta}_j^B)^2 - (y_j - \widehat{\eta}_j^A)^2\right\}\ \dot{\sim}\ \chi^2_{p-q}$$

  when model B is correct. This distribution is exact for linear models.
- ☐ If ϕ is unknown, it is replaced by an estimate. The large-sample properties of deviance differences outlined above still apply, though in small samples it may be better to replace the approximating χ^2 distribution by an F distribution with numerator degrees of freedom equal to the degrees of freedom for estimation of ϕ.

Model checking

- ☐ Need to assess whether a given model fits adequately, or needs to be modified.
- ☐ Two basic approaches:
  - overall tests either using generic statistic (e.g., chi-squared) or by **model expansion** (e.g., adding a term and testing for significance);
  - **regression diagnostics** for detecting a few possibly dodgy observations.
- ☐ Most widely used diagnostics in the linear model $y = X_{n\times p}\beta + \varepsilon$ are **residuals** $e_j = y_j - \widehat{y}_j$ and (much better) **standardized residuals**

$$r_j = \frac{y_j - \widehat{y}_j}{s(1-h_{jj})^{1/2}}, \quad j = 1, \ldots, n,$$

where the **leverage** h_{jj} is the jth diagonal element of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, and the **Cook statistic**

$$C_j = \frac{1}{ps^2}(\widehat{y} - \widehat{y}_{-j})^{\mathrm{T}}(\widehat{y} - \widehat{y}_{-j}) = \frac{r_j^2 h_{jj}}{p(1-h_{jj})},$$

which measures the effect of deleting the jth case (x_j, y_j) on the fitted model.

Diagnostics in general case

- ☐ Linear model ideas work as approximations (2nd order Taylor series, painful expansions).
- ☐ **Leverage** h_{jj} defined as jth diagonal element of

$$H = W^{1/2}X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W^{1/2},$$

depends in general on $\widehat{\beta}$, unlike in linear model.

- ☐ **Cook statistic** is change in deviance

$$C_j = 2p^{-1}\left\{\ell(\widehat{\beta}) - \ell(\widehat{\beta}_{-j})\right\} \doteq \frac{h_{jj}}{p(1-h_{jj})} r_{Pj}^2,$$

where $\widehat{\beta}_{-j}$ is MLE when jth case (x_j, y_j) is dropped, and r_{Pj} is **standardized Pearson residual** (see below).

- ☐ There are several types of residual (see next page).

Residuals in general case

- **Deviance residual**:
$$d_j = \text{sign}(\tilde\eta_j - \widehat\eta_j)[2\{\ell_j(\tilde\eta_j;\phi) - \ell_j(\widehat\eta_j;\phi)\}]^{1/2},$$
for which $\sum d_j^2 = D$ is deviance.
- **Pearson residual**: $u_j(\widehat\beta)/\sqrt{w_j(\widehat\beta)}$.
- Standardized versions
$$r_{Dj} = \frac{d_j}{(1-h_{jj})^{1/2}}, \quad r_{Pj} = \frac{u_j(\widehat\beta)}{\{w_j(\widehat\beta)(1-h_{jj})\}^{1/2}},$$
and (even better)
$$r_j^* = r_{Dj} + r_{Dj}^{-1}\log(r_{Pj}/r_{Dj}) \dot\sim N(0,1)$$
for many models.
- These all reduce to usual standardized residual for normal linear model.

Example

Example 33 (Gumbel linear model) *Give the components of the IWLS algorithm for fitting the linear model*
$$y_j = \beta_0 + \beta_1(x_j - \overline{x}) + \tau\varepsilon_j, \quad j = 1,\ldots,n,$$
with Gumbel errors having density function
$$f(y_j;\eta_j,\tau) = \tau^{-1}\exp\left\{-\frac{y_j-\eta_j}{\tau} - \exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\},$$
where $\tau > 0$ and $\eta_j = \beta_0 + \beta_1(x_j - \overline{x})$; this distribution is natural for maxima; note that τ^2 is not the variance.

Note to Example 33

- ☐ As the data are annual maxima, it is more appropriate to suppose that y_j has the Gumbel density

$$f(y_j;\eta_j,\tau) = \tau^{-1}\exp\left\{-\frac{y_j-\eta_j}{\tau} - \exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\}, \qquad (19)$$

where τ is a scale parameter and $\eta_j = \beta_0 + \beta_1(x_j - \overline{x})$; here we have replaced the γs with βs for continuity with the general discussion above.

- ☐ In this case

$$\ell_j(\eta_j,\tau) = -\log\tau - \frac{y_j-\eta_j}{\tau} - \exp\left(-\frac{y_j-\eta_j}{\tau}\right), \qquad (20)$$

and it is straightforward to establish that

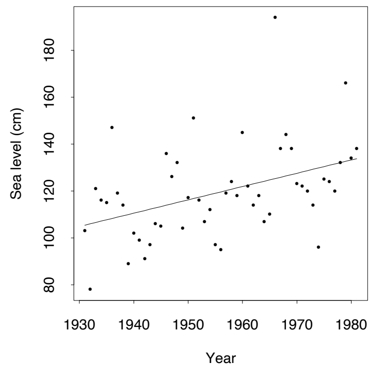
$$\frac{\partial \ell_j(\eta_j,\tau)}{\partial \eta_j} = \tau^{-1}\left\{1-\exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\}, \quad \mathrm{E}\left\{-\frac{\partial^2\ell_j(\eta_j,\tau)}{\partial\eta_j^2}\right\} = \tau^{-2},$$

that $\partial\eta/\partial\beta^{\mathrm{T}} = X$ is the $n\times 2$ matrix whose $j\mathrm{th}$ row is $(1, x_j-\overline{x})$, and $W = \tau^{-2}I_n$. Hence (17) becomes $\widehat{\beta} \doteq (X^{\mathrm{T}}X)^{-1}(X\beta+\tau^2 u)$, where the $j\mathrm{th}$ element of u is $\tau^{-1}[1-\exp\{-(y_j-\eta_j)/\tau\}]$.

- ☐ Here it is simplest to fix τ, to obtain $\widehat{\beta}$ by iterating (17) for each fixed value of τ, and then to repeat this over a range of values of τ, giving the profile log likelihood $\ell_{\mathrm{p}}(\tau)$ and hence confidence intervals for τ. Confidence intervals for β_0 and β_1 are obtained from the information matrix.
- ☐ With starting value chosen to be the least squares estimates of β, and with $\tau = 5$, 19 iterations of (17) were required to give estimates and a maximized log likelihood whose relative change was less than 10^{-6} between successive iterations. We then took $\tau = 5.5,\ldots,40$, using $\widehat{\beta}$ from the preceding iteration as starting-value for the next; in most cases just three iterations were needed. The left panel of Figure 2 shows a close-up of $\ell_{\mathrm{p}}(\tau)$; its maximum is at $\widehat{\tau} = 14.5$, and the 95% confidence interval for τ is $(11.9, 18.1)$. The maximum likelihood estimates of β_0 and β_1 are 111.4 and 0.563, with standard errors 2.14 and 0.137; these compare with standard errors 2.61 and 0.177 for the least squares estimates. There is some gain in precision in using the more appropriate model.

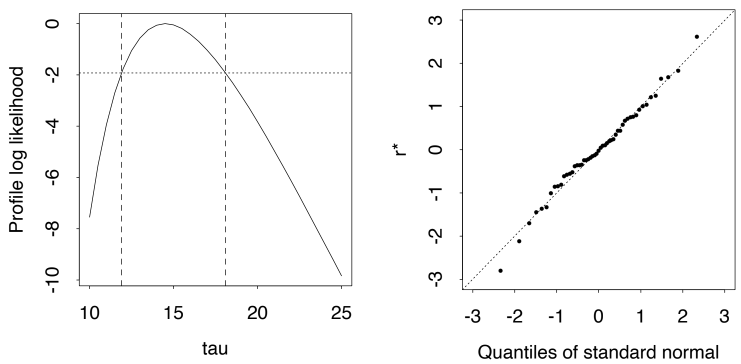
Venice data

Example 34 (Venice sea level data) *The figure below shows annual maximum sea levels in Venice, from 1931–1981. The very large value in 1966 is not an outlier. The fit of a Gumbel model to the data using IWLS gives MLEs (SEs)* $\widehat{\beta}_0 = 111.4\ (2.14)$ *(cm) and* $\widehat{\beta}_1 = 0.563\ (0.137)$ *(cm/year). The standard errors for LSEs are* $2.61,\ 0.177$*, larger than for MLEs with Gumbel model — gain in precision through using appropriate model.*



Venice data

Figure 2: Gumbel analysis of Venice data. Left panel: profile log likelihood $\ell_p(\tau) = \max_\beta \ell(\beta, \tau)$, with 95% confidence interval $(11.9, 18.1)$ (cm) for τ. Right panel: normal probability plot of residuals r_j^*.



Summary

- ☐ For regression problems with independent responses y_j dependent on parameters β through parameter $\eta_j = \eta(x_j; \beta)$, generalise least squares estimation to maximum likelihood estimation, using iterative weighted least squares algorithm: iterate to convergence

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z, \quad z = X\beta + W^{-1} u,$$

where

$$X_{n\times p} \equiv X(\beta) = \frac{\partial \eta}{\partial \beta^{\mathrm{T}}}, \quad u_{n\times 1} \equiv u(\eta) = \frac{\partial \ell}{\partial \eta}, \quad W_{n\times n} \equiv W(\eta) = -\mathrm{E}\left\{\frac{\partial^2 \ell}{\partial \eta \partial \eta^{\mathrm{T}}}\right\},$$

with ℓ the log likelihood for the data.

- ☐ Standard likelihood theory is used for confidence intervals and model comparison.
- ☐ Linear model diagnostics (residuals, leverage, Cook statistics, . . .) generalise to this setting.
- ☐ Next: generalized linear models (GLMs), wide class of models with exponential family-like response distributions.

Motivation

- Need to generalise linear model beyond normal responses, e.g. to data with $y \in \{0, 1, \ldots, m\}$, or $y \in \{0, 1, \ldots\}$, or $y > 0$.
- Consider **exponential family** response distributions (binomial, Poisson, ...), which have an elegant unifying theory, and encompass many possibilities (in addition to the normal)
- Basic idea is to build models such that

$$\mathrm{E}(y) = \mu, \quad g(\mu) = \eta = x^{\mathrm{T}}\beta,$$

  where g is a suitable function, and $y \sim$ exponential family (almost).
- **Warnings**:
  - **Don't** confuse Generalized Linear Model (GLM) with General Linear Model (GLM, in older books, the latter is $y = X\beta + \varepsilon$, with $\mathrm{cov}(\varepsilon) = \sigma^2 V$ not diagonal);
  - **Don't** write $y = \mu + \varepsilon$, since in a GLM the distribution of ε usually depends on μ.

Generalized linear model (GLM)

- Normal linear model has three key aspects:
  - structure for covariates: **linear predictor**, $\eta = x^{\mathrm{T}}\beta$;
  - response distribution: $y \sim N(\mu, \sigma^2)$;
  - linear relation $\eta = \mu$ between $\mu = \mathrm{E}(y)$ and η.
- GLM extends last two to
  - Y has density/mass function

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y; \phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Omega_\theta, \phi > 0, \tag{21}$$

    where
    - $\mathcal{Y}$ is the support of Y,
    - Ω_θ is the parameter space of valid values for $\theta \equiv \theta(\eta)$, and
    - the **dispersion parameter** ϕ is often known;
  - $\eta = g(\mu)$, where g is monotone **link function**
    - the **canonical link** function giving $\eta = \theta = b'^{-1}(\mu)$ has nice statistical properties;
    - but a range of link functions are possible for each distribution of Y.

Examples

Example 35 (GLM density) *Show that the moment-generating function of $f(y;\theta,\phi)$ is $M_Y(t)=\exp[\{b(\theta+t\phi)-b(\theta)\}/\phi]$, and deduce that*

$$\mathrm{E}(Y)=b'(\theta)=\mu,\quad \mathrm{var}(Y)=\phi b''(\theta)=\phi b''\{b'^{-1}(\mu)\}=\phi V(\mu);$$

the function $\mu\mapsto V(\mu)$ is known as the **variance function**.

Example 36 (Poisson distribution) *Write the Poisson mass function as a GLM density, and find its canonical link function.*

Example 37 (Normal distribution) *Write the normal density function as a GLM density, and find its canonical link function.*

Example 38 (Binomial distribution) *Write the binomial mass function as a GLM density, and find its canonical link function.*

Note to Example 35

- Suppose that Y has a continuous density; if not the argument below is the same, except that integrals are replaced by summations.
- Let $\Omega_\theta=\{\theta: b(\theta)<\infty\}$. Then

$$\begin{aligned} M_Y(t) &= \mathrm{E}\{\exp(tY)\}\\ &= \int e^{ty}\exp\left\{\frac{y\theta-b(\theta)}{\phi}+c(y;\phi)\right\}\,\mathrm{d}y\\ &= \int \exp\left\{\frac{y(\theta+t\phi)-b(\theta)}{\phi}+c(y;\phi)\right\}\,\mathrm{d}y. \end{aligned}$$

  If $\theta+t\phi\in\Omega_\theta$, then

$$\int \exp\left\{\frac{y(\theta+t\phi)-b(\theta+t\phi)}{\phi}+c(y;\phi)\right\}\,\mathrm{d}y=1,$$

  so

$$M_Y(t)=\mathrm{E}\{\exp(tY)\}=\exp\left[\{b(\theta+t\phi)-b(\theta)\}/\phi\right].$$

- Hence the cumulant-generating function of Y is

$$K_Y(t)=\log M_Y(t)=\{b(\theta+t\phi)-b(\theta)\}/\phi,$$

  and differentiating twice with respect to t and setting $t=0$ yields

$$\mathrm{E}(Y)=K_Y'(t)\big|_{t=0}=b'(\theta),\quad \mathrm{var}(Y)=K_Y''(t)\big|_{t=0}=\phi b''(\theta).$$

- One can show that $b(\theta)$ is strictly convex on Ω_θ. Thus $b'(\theta)$ is a monotonic increasing function of θ, so $b'^{-1}(\cdot)$ exists and is itself monotonic, so $V(\mu)=b''\{b'^{-1}(\mu)\}$ is well-defined.

Note to Example 36

The Poisson density may be written as

$$f(y;\mu) = \exp\left(y\log\mu - \mu - \log y!\right), \quad y = 0, 1, \ldots, \quad \mu > 0,$$

which has GLM form (21) with $\theta = \log\mu$, $b(\theta) = e^\theta$, $\phi = 1$, and $c(y;\phi) = -\log y!$. The mean of y is $\mu = b'(\theta) = e^\theta = \mu$, and its variance is $b''(\theta) = e^\theta = \mu$, so the variance function is linear: $V(\mu) = \mu$.

Note to Example 37

The normal density with mean μ and variance σ^2 may be written

$$f(y;\mu,\sigma^2) = \exp\left\{-\frac{(y^2 - 2y\mu + \mu^2)}{2\sigma^2} - \tfrac{1}{2}\log(2\pi\sigma^2)\right\},$$

so

$$\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \tfrac{1}{2}\theta^2, \quad c(y;\phi) = -\tfrac{1}{2\phi}y^2 - \tfrac{1}{2}\log(2\pi\phi).$$

As the first and second derivatives of $b(\theta)$ are θ and 1, we have $V(\mu) = 1$; the variance function is constant.

Note to Example 38

We write the binomial density

$$f(r;\pi) = \binom{m}{r}\pi^r(1-\pi)^{m-r}, \quad 0 < \pi < 1, \quad r = 0, \ldots, m,$$

in the form

$$\exp\left[m\left\{\frac{r}{m}\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} + \log\binom{m}{r}\right],$$

so

$$y = \frac{r}{m}, \ \phi = \frac{1}{m}, \ \theta = \log\left(\frac{\pi}{1-\pi}\right), \ b(\theta) = \log(1 + e^\theta), \ c(y;\phi) = \log\binom{m}{r}.$$

The mean and variance of y are

$$\mu = b'(\theta) = \frac{e^\theta}{1 + e^\theta}, \quad \phi b''(\theta) = \frac{e^\theta}{m(1 + e^\theta)^2};$$

the variance function is $V(\mu) = \mu(1-\mu)$.

Estimation of β

Example 39 (IWLS algorithm) *Find the components of the IWLS algorithm for a GLM.*

- ☐ If canonical link is used then $\theta_j = x_j^{\mathrm{T}}\beta$, so if ϕ is known, then

$$\begin{aligned}\ell(\beta) &= \sum_{j=1}^{n}\left\{\frac{y_j x_j^{\mathrm{T}}\beta - b(x_j^{\mathrm{T}}\beta)}{\phi} + c(y_j;\phi)\right\}\\ &= \{y^{\mathrm{T}}X\beta - K(\beta)\}/\phi + C(y;\phi),\end{aligned}$$

  say, which in terms of β is a linear exponential family with
  - **canonical parameter** $\beta_{p\times 1}$
  - **canonical statistic** $(X^{\mathrm{T}}y)_{p\times 1}$,

  and many nice properties then hold.
- ☐ If X is full rank, then $\ell(\beta)$ is (almost always) strictly concave and has a unique maximum in terms of β.
- ☐ Problem: the maximum may be at infinity in certain (rare) cases—this can arise with binomial responses: beware of $\widehat{\theta}_r \approx \pm 36$.

Note to Example 39

☐ To compute the quantities needed for the IWLS step $\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W(X\beta + W^{-1}u)$, we need

$$X_{n\times p} = \frac{\partial\eta}{\partial\beta^{\mathrm{T}}}, \quad W_{n\times n} = \mathrm{diag}\{\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)\}, \quad u_{n\times 1} = \{\partial\ell_j/\partial\eta_j\},$$

where (with ϕ_j instead of ϕ for generality, see the next slide),

$$\ell_j(\beta) = \left\{\frac{y_j\theta_j - b(\theta_j)}{\phi_j} + c(y_j;\phi_j)\right\}, \quad b'(\theta_j) = \mu_j, \quad \eta_j = g(\mu_j) = x_j^{\mathrm{T}}\beta.$$

☐ First note that $\partial\eta_j/\partial\beta_r = x_{jr}$, so $X = \partial\eta/\partial\beta^{\mathrm{T}}$ is just a matrix of constants.

☐ We need the first and second derivatives of ℓ_j with respect to η_j, so we write

$$\frac{\partial\ell_j}{\partial\eta_j} = \frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\theta_j}{\partial\mu_j}\frac{\partial\ell_j}{\partial\theta_j},$$

with

$$\frac{\partial\eta_j}{\partial\mu_j} = g'(\mu_j), \quad \frac{\partial\mu_j}{\partial\theta_j} = b''(\theta_j) = V(\mu_j), \quad \frac{\partial\ell_j}{\partial\theta_j} = \frac{y_j - b'(\theta_j)}{\phi_j},$$

which yields

$$u_j = \frac{\partial\ell_j}{\partial\eta_j} = \frac{y_j - b(\theta_j)}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{y_j - \mu_j}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{A(\theta_j)}{B(\theta_j)},$$

say, where $\mathrm{E}(A) = 0$. For the second derivative, we note that

$$\frac{\partial^2\ell_j}{\partial\eta_j^2} = \frac{\partial}{\partial\eta_j}\frac{\partial\ell_j}{\partial\eta_j} = \left(\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\theta_j}{\partial\mu_j}\frac{\partial}{\partial\theta_j}\right)\frac{\partial\ell_j}{\partial\eta_j} = \frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\theta_j}{\partial\mu_j}\left\{\frac{A'(\theta_j)}{B(\theta_j)} - \frac{A(\theta_j)B'(\theta_j)}{B(\theta_j)^2}\right\},$$

and on noting that $B(\theta_j)$ is non-random and $A'(\theta_j) = -b''(\theta_j) = -V(\mu_j)$, we obtain

$$w_j = \mathrm{E}\left(-\frac{\partial^2\ell_j}{\partial\eta_j^2}\right) = \frac{1}{g'(\mu_j)}\frac{1}{V(\mu_j)}\frac{V(\mu_j)}{g'(\mu_j)\phi_j V(\mu_j)} = \frac{1}{g'(\mu_j)^2\phi_j V(\mu_j)}.$$

Note to Example 39, part II

- From above we see that the components of the score statistic $u(\beta)$ and the weight matrix $W(\beta)$ may be expressed in terms of components μ_j of the mean vector μ as

$$\begin{aligned} u_j &= \frac{\partial\theta_j}{\partial\eta_j}\frac{\partial\ell_j(\theta_j)}{\partial\theta_j} = \frac{y_j-\mu_j}{g'(\mu_j)\phi_j V(\mu_j)}, \\ w_j &= \left(\frac{\partial\theta_j}{\partial\eta_j}\right)^2\frac{\partial^2\ell_j(\theta_j)}{\partial\theta_j^2} = \frac{1}{g'(\mu_j)^2\phi_j V(\mu_j)}, \end{aligned} \qquad (22)$$

  where $g'(\mu_j) = dg(\mu_j)/d\mu_j$. Thus $\widehat{\beta}$ is obtained by iterative weighted least squares regression of response

$$z = X\beta + g'(\mu)(y-\mu) = \eta + g'(\mu)(y-\mu)$$

  on the columns of X using weights (22).
- By using y as an initial value for μ and $g(y)$ as an initial value for $\eta = X\beta$, we avoid needing an initial value for β.
- It may be necessary to modify y slightly for this initial step. For example if we use the log link for Poisson data, and some y_j equal zero, then we may need to replace them with some small positive value to avoid taking $\log 0$ for some components of the initial $\eta = \log y$.

Estimation of ϕ

- When ϕ unknown, it is often replaced by $\phi_j = \phi a_j$, with known a_j and a_j^{-1} treated as a weight. Then we replace the scaled deviance by the **deviance** ϕD.
- If the model is correct and ϕ is known, then **Pearson's statistic**

$$P = \frac{1}{\phi}\sum_{j=1}^{n}\frac{(y_j-\widehat{\mu}_j)^2}{a_j V(\widehat{\mu}_j)} \dot{\sim} \chi^2_{n-p},$$

  analogously to the sum of squares in a linear model, with $\mathrm{E}(P) \doteq n-p$.
- The MLE of ϕ can be badly behaved, so usually we prefer the method of moments estimator

$$\widehat{\phi} = \frac{1}{n-p}\sum_{j=1}^{n}(y_j-\widehat{\mu}_j)^2/\{a_j V(\widehat{\mu}_j)\},$$

  which is obtained by solving the equation $P = n-p$, based on noting that $\mathrm{E}(\chi^2_{n-p}) = n-p$.
- If the data are sparse (e.g., many small binomial or Poisson counts), then standard asymptotic results are suspect.

Example: Jacamar data

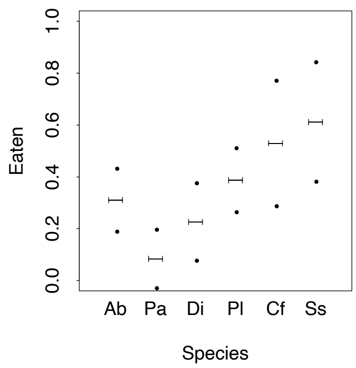
Table 4: Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

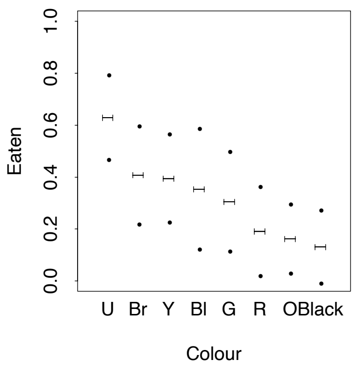
	Aphrissa boisduvalli N/S/E	*Phoebis argante* N/S/E	*Dryas iulia* N/S/E	*Pierella luna* N/S/E	*Consul fabius* N/S/E	*Siproeta stelenes*† N/S/E
Unpainted	0/0/14	6/1/0	1/0/2	4/1/5	0/0/0	0/0/1
Brown	7/1/2	2/1/0	1/0/1	2/2/4	0/0/3	0/0/1
Yellow	7/2/1	4/0/2	5/0/1	2/0/5	0/0/1	0/0/3
Blue	6/0/0	0/0/0	0/0/1	4/0/3	0/0/1	0/1/1
Green	3/0/1	1/1/0	5/0/0	6/0/2	0/0/1	0/0/3
Red	4/0/0	0/0/0	6/0/0	4/0/2	0/0/1	3/0/1
Orange	4/2/0	6/0/0	4/1/1	7/0/1	0/0/2	1/1/1
Black	4/0/0	0/0/0	1/0/1	4/2/2	7/1/0	0/1/0

† includes *Philaethria dido* also.

Jacamar data

Figure 3: Proportion of butterflies eaten ($\pm 2SE$) for different species and wing colour.





Jacamar data

- ☐ How does a bird respond to the species s and wing colour c of its prey?
- ☐ Response has 3 (ordered) categories: not attacked (N), attacked but then rejected (S), attacked and eaten (E)
- ☐ The data form an 8×6 layout, with a 3-category response in each cell, total m_{cs}
- ☐ Assume that the number in category E (response) is binomial:

$$R_{cs} \sim B(m_{cs}, \pi_{cs}), \quad c = 1, \ldots, 8, s = 1, \ldots, 6,$$

  where c is colour and s is species, with probability that bird attacks and eats butterfly is

$$\pi_{cs} = \frac{\exp(\alpha_c + \gamma_s)}{1 + \exp(\alpha_c + \gamma_s)}, \quad c = 1, \ldots, 8, s = 1, \ldots, 6,$$

  so
  - large α_c corresponds to colours that the jacamar likes to eat,
  - large γ_s corresponds to species that it likes.
- ☐ This is a GLM with response $y_{cs} = r_{cs}/m_{cs}$, $\mathrm{E}(y_{cs}) = \pi_{cs}$, and canonical (logit) link function

$$\eta = \log\{\pi/(1-\pi)\}, \quad \eta_{cs} = \alpha_c + \gamma_s.$$

Jacamar data: Analysis of deviance

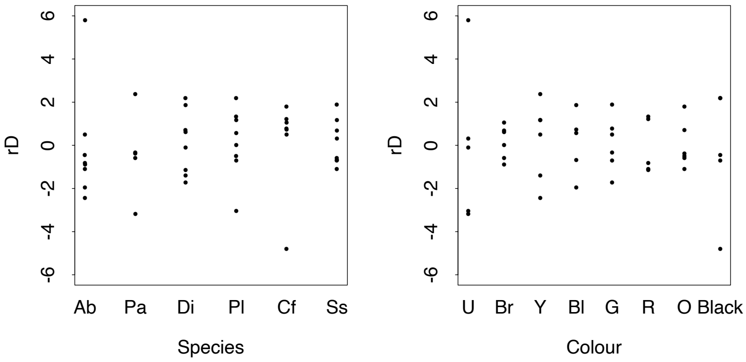
Table 5: Deviances and analysis of deviance for models fitted to jacamar data. The lower part shows results for the reduced data, without two outliers.

	Full data		Without outliers	
Terms	df	Deviance	df	Deviance
1	43	134.24	35	73.68
1+Species	38	114.59	31	46.04
1+Colour	36	108.46	28	63.20
1+Species+Colour	31	67.28	24	28.02

Terms	df	Deviance reduction	Terms	df	Deviance reduction
Species (unadj. for Colour)	5	19.64	Species (adj. for Colour)	5	41.18
Colour (adj. for Species)	7	47.31	Colour (unadj. for Species)	7	25.78
Species (unadj. for Colour)	4	27.63	Species (adj. for Colour)	4	35.18
Colour (adj. for Species)	7	18.03	Colour (unadj. for Species)	7	10.48

Jacamar data: Residuals

Figure 4: Standardized deviance residuals r_D for binomial two-way layout fitted to jacamar data.



Jacamar data: Parameter estimates

Table 6: Estimated parameters and standard errors for the jacamar data, without 2 outliers.

Aphrissa boisduvalli	*Phoebis argante*	*Dryas iulia*	*Pierella luna*	*Consul fabius*	*Siproeta stelenes*
–1.99 (0.79)	–2.22 (0.85)	–0.56 (0.67)	0.16 (0.54)	—	1.50 (0.78)

Brown	Yellow	Blue	Green	Red	Orange	Black
0.16 (0.73)	0.33 (0.68)	–0.53 (0.81)	–0.83 (0.75)	–1.93 (0.88)	–1.94 (0.85)	–1.26 (0.86)

- ☐ Interpretation
- ☐ Residual deviance: 28.02, with 24 df
- ☐ Pearson statistic: 25.58, with 24 df
- ☐ Standardized residuals in range -2.03 to 1.96: OK.

Example: Chimpanzee data

Table 7: Times in minutes taken by four chimpanzees to learn ten words.

Chimpanzee	Word									
	1	2	3	4	5	6	7	8	9	10
1	178	60	177	36	225	345	40	2	287	14
2	78	14	80	15	10	115	10	12	129	80
3	99	18	20	25	15	54	25	10	476	55
4	297	20	195	18	24	420	40	15	372	190

- ☐ A two-way layout.
- ☐ Times vary from 2 to 476 minutes — need transformation (e.g., logarithm) if use linear model.

Chimpanzee data

- ☐ How does learning time depend on word w and chimp c?
- ☐ Response is continuous and positive, so we try fitting the gamma distribution with mean μ and shape parameter ν, i.e.,

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)} y^{\nu-1} \left(\frac{\nu}{\mu}\right)^{\nu} \exp(-\nu y/\mu), \quad y > 0, \quad \nu, \mu > 0,$$

  so dispersion parameter is $\phi = 1/\nu$ ($\phi = \nu = 1$ for exponential).
- ☐ Possible link functions:

$$\eta = \log \mu, \text{ (log, most common)}, \quad \eta = 1/\mu, \text{ (reciprocal, canonical)}$$

- ☐ Linear model structure:

$$\eta_{cw} = \alpha_c + \gamma_w, \quad c = 1, \ldots, 4, w = 1, \ldots, 10,$$

  but the interpretation of the α_c and γ_w will depend on the link function.
- ☐ With the log link, the deviances for models 1, 1+Chimp, 1+Word, and 1+Chimp+Word are 60.38, 53.43, 21.19, and 14.97. How many df are there for each model?

Chimpanzee data: Analysis of deviance

Table 8: Analysis of deviance for models fitted to chimpanzee data.

Term	df	Deviance reduction	Term	df	Deviance reduction
Chimp (unadj. for Word)	3	6.95	Chimp (adj. for Word)	3	6.22
Word (adj. for Chimp)	9	38.46	Word (unadj. for Chimp)	9	39.19

- ☐ Method of moments estimate is $\widehat{\phi} = 0.432$, so $\widehat{\nu} = 1/\widehat{\phi} = 2.31$.
- ☐ Use F tests to assess effects of Word and Chimp, for example obtaining

$$\frac{6.22/3}{0.423} = 4.78 \dot{\sim} F_{3,27}$$

  if there is no difference between the chimps. What is the corresponding statistic for testing differences between words?
- ☐ Residuals suggest that this model, or one with the inverse link, are both adequate, and both are better than fitting a normal linear model to the log times.

Summary

- ☐ Generalized linear models extend the classical linear model in two ways:
  - the response distribution is (almost) exponential family, so includes binomial, Poisson, gamma and other distributions in addition to the normal;
  - the relation between the linear predictor $\eta = x^{\mathrm{T}}\beta$ and the mean μ is determined by a wide range of possible link functions.
- ☐ Canonical link functions give particularly simple models and are widely used.
- ☐ Estimates of β are obtained by IWLS, which has a simple form, with no need for initial values.
- ☐ A simple estimate of the dispersion parameter ϕ is available using the method of moments.
- ☐ Models are compared using the analysis of deviance, which generalises the analysis of variance in the classical linear model.
- ☐ Standard likelihood theory results are used for inference (standard errors, confidence intervals, etc.)
- ☐ Standard diagnostics (residuals, . . .) extend in a natural way to this setting.

Binary response

- Response Y has Bernoulli distribution with

$$\mathrm{P}(Y=1)=\pi, \quad \mathrm{P}(Y=0)=1-\pi, \quad 0<\pi<1.$$

and $\mathrm{E}(Y)=\mu=\pi$, $\mathrm{var}(Y)=\pi(1-\pi)$.

- Linear link function $\pi=\eta=x^{\mathrm{T}}\beta$ can give $\pi\notin[0,1]$, so not usually a good idea.
- Y can be interpreted in terms of a hidden variable/tolerance distribution: let $Z=x^{\mathrm{T}}\gamma+\sigma\varepsilon$, where $\varepsilon\sim F$. Set $Y=I(Z>0)$, and note that

$$\pi=\mathrm{P}(Y=1)=\mathrm{P}(x^{\mathrm{T}}\gamma+\sigma\varepsilon>0)=\mathrm{P}(\varepsilon>-x^{\mathrm{T}}\gamma/\sigma)=1-F(-x^{\mathrm{T}}\beta),$$

say. Note that $\beta=\gamma/\sigma$ is estimable, but γ and σ are not.

- The corresponding link function is given by

$$\eta=x^{\mathrm{T}}\beta=-F^{-1}(1-\pi)=g(\pi),$$

so different choices of F yield different possible link functions.

Link functions

Tolerance distributions and corresponding link functions for binary data.

Distribution F		Link function	
Logistic	$e^u/(1+e^u)$	Logit	$\eta=\log\{\pi/(1-\pi)\}$
Normal	$\Phi(u)$	Probit	$\eta=\Phi^{-1}(\pi)$
Log Weibull	$1-\exp(-\exp(u)\}$	Log-log	$\eta=-\log\{-\log(\pi)\}$
Gumbel	$\exp\{-\exp(-u)\}$	Complementary log-log	$\eta=\log\{-\log(1-\pi)\}$

- The logit and probit links are symmetric.
- Logit (canonical link) is usual choice, good for medical studies (later), with nice interpretation, but the probit is very similar to it and may be preferred in some cases, for its relation to the normal distribution.
- The log-log and complementary log-log links are asymmetric.

Logistic regression

- Commonest choice of link function for proportion data is the **logit**, which gives

$$P(Y=1)=\pi=\frac{\exp(x^{\mathrm{T}}\beta)}{1+\exp(x^{\mathrm{T}}\beta)},\quad P(Y=0)=1-\pi=\frac{1}{1+\exp(x^{\mathrm{T}}\beta)},$$

leading to a linear model for the **log odds** of success,

$$\log\left\{\frac{P(Y=1)}{P(Y=0)}\right\}=\log\left(\frac{\pi}{1-\pi}\right)=x^{\mathrm{T}}\beta,\quad \beta\in\mathbb{R}^p.$$

- The likelihood for β based on independent responses $y_1,\ldots,y_n$ with covariate vectors $x_1,\ldots,x_n$ and corresponding probabilities $\pi_1,\ldots,\pi_n$ is

$$L(\beta)=\prod_{j=1}^{n}\pi_j^{y_j}(1-\pi_j)^{1-y_j}=\cdots=\frac{\exp\left(\sum_{j=1}^{n}y_jx_j^{\mathrm{T}}\beta\right)}{\prod_{j=1}^{n}\left\{1+\exp\left(x_j^{\mathrm{T}}\beta\right)\right\}},$$

which is a regular exponential family with $s(y)=X^{\mathrm{T}}y$ and log likelihood

$$\ell(\beta)=(X^{\mathrm{T}}y)^{\mathrm{T}}\beta-\sum_{j=1}^{n}\log\left\{1+\exp\left(x_j^{\mathrm{T}}\beta\right)\right\},\quad \beta\in\mathbb{R}^p,$$

known as the **logistic regression model**.

Nodal involvement data

Data on nodal involvement: 53 patients with prostate cancer have nodal involvement (r), with five binary covariates age, stage, etc.

m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	
1	1	0	0	1	0	1
1	0	0	0	0	1	1
1	0	0	0	0	1	0

Deviances for nodal involvement models

Scaled deviances D for 32 logistic regression models for nodal involvement data. $+$ denotes a term included in the model.

age	st	gr	xr	ac	df	D	age	st	gr	xr	ac	df	D
					52	40.71	+	+	+			49	29.76
+					51	39.32	+	+		+		49	23.67
	+				51	33.01	+	+			+	49	25.54
		+			51	35.13	+		+	+		49	27.50
			+		51	31.39	+		+		+	49	26.70
				+	51	33.17	+			+	+	49	24.92
+	+				50	30.90		+	+	+		49	23.98
+		+			50	34.54		+	+		+	49	23.62
+			+		50	30.48		+		+	+	49	19.64
+				+	50	32.67			+	+	+	49	21.28
	+	+			50	31.00	+	+	+	+		48	23.12
	+		+		50	24.92	+	+	+		+	48	23.38
	+			+	50	26.37	+	+		+	+	48	19.22
		+	+		50	27.91	+		+	+	+	48	21.27
		+		+	50	26.72		+	+	+	+	48	18.22
			+	+	50	25.25	+	+	+	+	+	47	18.07

Model selection

- We have 32 competing models, and would like to select the 'best', or a few 'near-best'.
- In general we have 2^p models, so automatic selection of some sort is helpful.
- Could use likelihood ratio tests (differences of deviances) to compare competing models, but this involves many correlated tests, so may lead to spurious results.
- Usually minimise an information criterion, which accounts for the number of parameters in each model, such as

$$\mathrm{AIC} \equiv D + 2p, \quad \mathrm{BIC} \equiv D + p \log n,$$

  where D is the deviance.
- Recall their properties, with p fixed and as $n \to \infty$:
  - AIC tends to overfit, i.e., it has a positive probability of choosing a model that is too complex,;
  - BIC applies a stronger penalty, so *if the true model is among those fitted*, it will choose it with probability one;
  - BIC usually yields less complex models than AIC, but they may predict less well.
- There are many other information criteria, but these are most used in practice.

Example: Nodal involvement

☐ Model with lowest AIC has stage, xray, acid:

$$x^{\mathrm{T}}\widehat{\beta} = -3.05 + 1.65 I_{\textsf{stage}} + 1.91 I_{\textsf{xray}} + 1.64 I_{\textsf{acid}},$$

where $I_{\textsf{stage}} = 1$ indicates that stage takes its higher level, etc.

☐ Interpretation of this model:

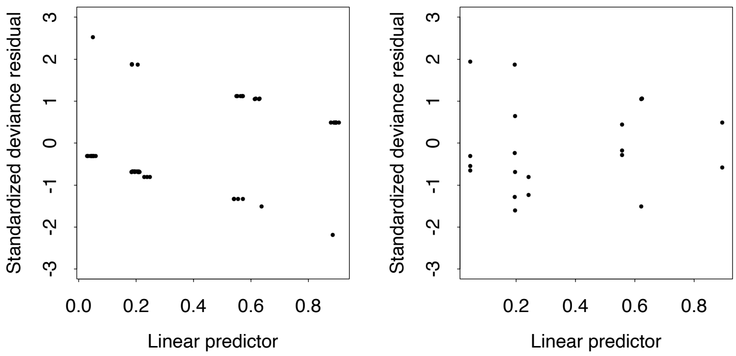
- for an individual with stage, xray and acid at their lowest levels, the fitted probability of nodal involvement is $e^{-3.05}/(1 + e^{-3.05}) \doteq 0.045$ (though there are no such people in the data, so this involves extrapolation);
- for someone with only $I_{\textsf{stage}} = 1$, the odds of nodal involvement are $e^{-3.05+1.65} = e^{-1.4} \doteq 0.25$, a probability of 0.2;
- for someone with $I_{\textsf{stage}} = I_{\textsf{xray}} = I_{\textsf{acid}} = 1$, the odds of nodal involvement are $e^{-3.05+1.65+1.91+1.64} \doteq 8.6$, a probability of 0.9;

☐ Problems with interpretation of residual deviance of 19.64: how many df? — can amalgamate independent binary responses with same covariates.

☐ Likewise problems with residuals . . .

Nodal involvement residuals

Figure 5: Standardized deviance residuals for nodal involvement data, for ungrouped responses (left) and grouped responses (right).



Summary

- ☐ Proportion data are often modelled using the Bernoulli/binomial response distributions.
- ☐ Link functions (logit, probit, . . .) have interpretations in terms of underlying continuous variables that have been dichotomized.
- ☐ The canonical and most commonly-used link is the logit, and fitting using this yields logistic regression, in which
  - the canonical parameter is the log odds;
  - classical data structures (e.g., the 2×2 table) have nice interpretations.
- ☐ The deviance can be used to compare models (so can AIC, BIC, . . .), but using its absolute value to assess fit can be dangerous (exercise).
- ☐ Residuals for binary data are not very informative.

Types of count data

- $y \in \{0, 1, 2, \ldots\}$, perhaps with upper bound m, depending on sampling scheme:
  - counts, with no fixed total;
  - m individuals, subdivided into various categories:
    - **nominal response**—unordered categories (gender, nationality, ...)
    - **ordinal response**—ordered categories (pain level, spiciness of curry, ...)
- Simplest models:
  - single unbounded response, or Poisson approximation to binomial, takes $Y \sim \text{Pois}(\mu)$;
  - group of responses $(Y_1, \ldots, Y_d)$ with fixed total $\sum Y_j = m$ has multinomial distribution, probabilities $(\pi_1, \ldots, \pi_d)$ and denominator m.
- Previous examples:
  - Doll and Hill data on smoking had response y Poisson with $\mu = T\lambda(x; \beta)$;
  - Jacamar data had ordinal (?) response N/S/E with total N+S+E fixed—multinomial with $d = 3$

Poisson and multinomial distributions

- $Y \sim \text{Pois}(\mu)$ implies that

$$f(y; \mu) = \frac{\mu^y}{y!} e^{-\mu}, \quad y = 0, 1, 2, \ldots, \quad \mu > 0.$$

- Exponential family with natural parameter $\theta = \log \mu$, GLM with canonical logarithmic link, $x^{\mathrm{T}}\beta = \eta = \log \mu$.
- If Y is number of events in Poisson process of rate λ observed for period of length T, then $\mu = \lambda T$ and we set $\eta = x^{\mathrm{T}}\beta + \log T$
  - **offset** $\log T$ is fixed part of linear predictor η
- If $Y_r \overset{\text{ind}}{\sim} \text{Pois}(\mu_r)$, $r = 1, \ldots, d$, then the joint distribution of $Y_1, \ldots, Y_d$ given $Y_1 + \cdots + Y_d = m$ is **multinomial,** with denominator m, and probabilities

$$\pi_1 = \frac{\mu_1}{\sum_{r=1}^d \mu_r}, \quad \ldots, \quad \pi_d = \frac{\mu_d}{\sum_{r=1}^d \mu_r}.$$

- If $(Y_1, \ldots, Y_d) \sim \text{Mult}(m; \pi_1, \ldots, \pi_d)$, then marginal and conditional distributions, e.g., of

$$(Y_1 + Y_2, Y_3 + Y_4 + Y_5, Y_6, \ldots, Y_d), \quad (Y_1, Y_2, Y_4) \mid (Y_3, Y_5, \ldots, Y_d),$$

are also multinomial.

Log-linear and logistic regressions

- Special case: if $d = 2$, then

$$Y_2 \mid Y_1 + Y_2 = m \quad \sim \quad B\left(m, \pi = \frac{\mu_2}{\mu_1 + \mu_2}\right)$$

- If $\mu_1 = \exp(\gamma + x_1^{\mathrm{T}}\beta)$, $\mu_2 = \exp(\gamma + x_2^{\mathrm{T}}\beta)$, then

$$\pi = \frac{\exp(\gamma + x_2^{\mathrm{T}}\beta)}{\exp(\gamma + x_1^{\mathrm{T}}\beta) + \exp(\gamma + x_2^{\mathrm{T}}\beta)} = \frac{\exp\{(x_2 - x_1)^{\mathrm{T}}\beta\}}{1 + \exp\{(x_2 - x_1)^{\mathrm{T}}\beta\}},$$

  which corresponds to a logistic regression model for Y_2 with denominator m and probability π.
- Can estimate β using log linear model or logistic model—but can't estimate γ from logistic model.

Poisson Regression slide 274

Premier League data

```
> soccer
    month day year       team1       team2 score1 score2
1     Aug  19 2000    Charlton ManchesterC      4      0
2     Aug  19 2000     Chelsea     WestHam      4      2
3     Aug  19 2000    Coventry   Middlesbr      1      3
4     Aug  19 2000       Derby Southampton      2      2
5     Aug  19 2000       Leeds     Everton      2      0
6     Aug  19 2000   Leicester  AstonVilla      0      0
7     Aug  19 2000   Liverpool    Bradford      1      0
8     Aug  19 2000  Sunderland     Arsenal      1      0
9     Aug  19 2000   Tottenham     Ipswich      3      1
10    Aug  20 2000 ManchesterU   Newcastle      2      0
11    Aug  21 2000     Arsenal   Liverpool      2      0
12    Aug  22 2000    Bradford     Chelsea      2      0
13    Aug  22 2000     Ipswich ManchesterU      1      1
14    Aug  22 2000   Middlesbr   Tottenham      1      1
15    Aug  23 2000     Everton    Charlton      3      0
16    Aug  23 2000 ManchesterC  Sunderland      4      2
17    Aug  23 2000   Newcastle       Derby      3      2
18    Aug  23 2000 Southampton    Coventry      1      2
19    Aug  23 2000     WestHam   Leicester      0      1
20    Aug  26 2000     Arsenal    Charlton      5      3
 ...
```

Premier League data

- ☐ 380 soccer matches in English Premier League in 2000–2001 season.
- ☐ Data: home score y_{ij}^h and away score y_{ij}^a when team i is at home to team j, for $i, j, = 1, \ldots, 20$, $i \neq j$.
- ☐ Treat these as Poisson counts with means

$$\mu_{ij}^h = \exp(\Delta + \alpha_i - \beta_j), \quad \mu_{ij}^a = \exp(\alpha_j - \beta_i)$$

where

  - Δ represents the home advantage;
  - α_i and β_i represent the offensive and defensive strengths of team i.
- ☐ Two possibilities for fitting:
  - Poisson GLM, with 39 parameters;
  - binomial GLM, with 20 parameters.

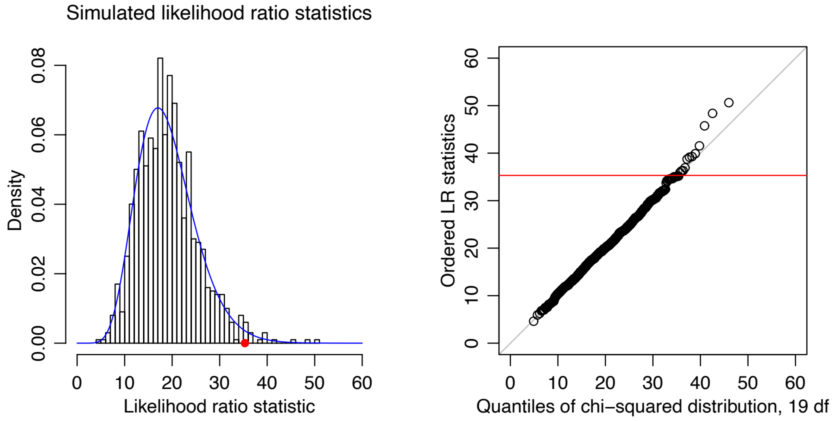
Premier League data: Analysis of deviance

Poisson model			Binomial model		
Terms	df	Deviance reduction	Terms	df	Deviance reduction
Home	1	33.58	Home	1	33.58
Defence	19	39.21	Team	19	79.63
Offence	19	58.85			
Residual	720	801.08	Residual	332	410.65

- ☐ There's a strong effect of playing at home, and lots of evidence of differences among the teams—more in offence than defence.
- ☐ Both residual deviances are a little large, but since the counts are small, we don't expect the large-sample χ^2 distribution to apply well to the residual deviance.
- ☐ Simulations from the fitted model suggest that the residual deviances are not unusually large, so there's no evidence of a lack of fit.

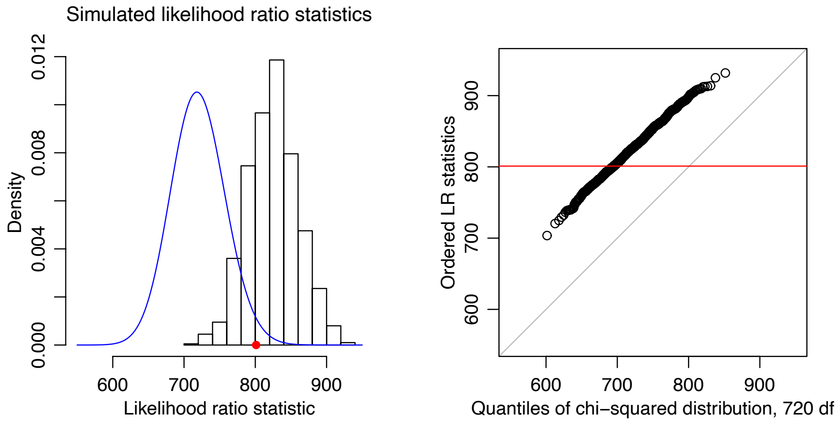
Premier League data: Null deviance for defence effect

Defence effect deviance (in red) for the Poisson model is large(ish) relative to χ^2_{19} distribution, but the asymptotics seem OK, based on simulations from a model without this effect (i.e., `Home + Offence`). It seems we can trust asymptotic distributions for differences of deviances, even though the counts are small.



Premier League data: Residual deviance

Residual deviance of 801 (in red) for the Poisson model seems large(ish) relative to χ^2_{720} distribution, but the asymptotics are suspect because most of the counts are small. Comparison of observed deviance with χ^2_{720} distribution shows that 801 is in fact somewhat smaller than average for datasets simulated from the fitted model.



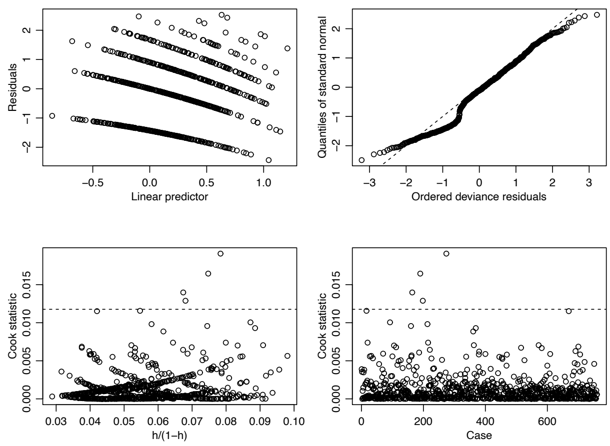
Premier League data: Estimates

	Overall (δ)	Offensive (α)	Defensive (β)
Manchester United	0.39	0.22	0.15
Liverpool	0.13	0.12	-0.08
Arsenal	—	0.04	—
Chelsea	-0.09	0.08	-0.22
Leeds	-0.10	0.02	-0.17
Ipswich	-0.16	-0.10	-0.13
Sunderland	-0.33	-0.31	-0.10
Aston Villa	-0.48	-0.31	-0.15
West Ham	-0.53	-0.33	-0.30
Middlesborough	-0.53	-0.35	-0.17
Charlton	-0.55	-0.21	-0.43
Tottenham	-0.58	-0.28	-0.38
Newcastle	-0.59	-0.35	-0.30
Southampton	-0.60	-0.45	-0.25
Everton	-0.75	-0.32	-0.46
Leicester	-0.77	-0.47	-0.31
Manchester City	-0.90	-0.40	-0.56
Coventry	-0.93	-0.53	-0.52
Derby	-0.93	-0.51	-0.45
Bradford	-1.29	-0.71	-0.62
SEs	0.29	0.20	0.20

Home advantage: $\widehat{\Delta} = 0.37$ (0.07), $\exp(\widehat{\Delta}) = 1.45$.

Premier League data: Assessment of fit

Diagnostic plots for fitted model: residuals against $\widehat{\eta}$ (top left); normal QQ-plot of residuals (top right); Cook statistic C_j against leverage ratio $h_j/(1-h_j)$ (lower left); Cook statistic C_j against case number (lower right).



Sampling schemes

- A **contingency table** contains individuals (sampling units) cross-classified by various categorical variables.
  - Example: the jacamar data cross-classify butterflies by

$$6 \text{ species } \times \text{ 8 colours } \times \text{ 3 fates}$$

    for a total of 144 categories, each with its number of butterflies $0, 1, \ldots, 14$.
- The sampling scheme underlying a table may fix certain totals. Suppose a pollster wants to find out how people will vote. She might
  - wait in the street for a morning, and get opinions from those people willing to talk to her;
  - wait until she has the views of a fixed number, say m, of people;
  - wait until she has the views of fixed numbers of men and women.

Example 40 *Find the likelihoods for each of these sampling schemes, under (unrealistic!) assumptions of independence of voters.*

Note to Example 40

- An $R \times C$ table arises by randomly sampling a population over a fixed period and then classifying the resulting individuals.
- In the first scheme there are no constraints on the row and column totals, and a simple model is that the count in the (r, c) cell, y_{rc}, has a Poisson distribution with mean μ_{rc}. The resulting likelihood is

$$\prod_{r,c} \left\{ \frac{\mu_{rc}^{y_{rc}}}{y_{rc}!} e^{-\mu_{rc}} \right\};$$

  this is simply the Poisson likelihood for the counts in the RC groups.
- The pollster may set out with the intention of interviewing a fixed number m of individuals, stopping only when $\sum_{rc} y_{rc} = m$. In this case the data are multinomially distributed, with likelihood

$$\frac{m!}{\prod_{r,c} y_{rc}!} \prod_{r,c} \pi_{rc}^{y_{rc}}, \quad \sum_{r,c} \pi_{rc} = 1,$$

  with $\pi_{rc} = \mu_{rc} / \sum_{s,t} \mu_{st}$ the probability of falling into the (r, c) cell.
- A third scheme is to interview fixed numbers of men and of women, thus fixing the row totals $m_r = \sum_c y_{rc}$ in advance. In effect this treats the row categories as subpopulations, and the column categories as the response. This yields independent multinomial distributions for each row, and product multinomial likelihood

$$\prod_r \left\{ \frac{m_r!}{\prod_c y_{rc}!} \prod_c \pi_{rc}^{y_{rc}} \right\}, \quad \sum_c \pi_{1c} = \cdots = \sum_c \pi_{Rc} = 1,$$

  in which $\pi_{rc} = \mu_{rc} / \sum_t \mu_{rt}$.

Contingency tables and Poisson response models

- ☐ Multinomial models can be fitted using Poisson errors, provided the appropriate baseline terms are always included in the linear predictor.
- ☐ Write the data as two-way layout, with C columns and R rows with fixed totals (e.g., $6 \times 8 = 48$ rows each with 3 columns for the jacamar data).
- ☐ Consider Poisson model with means $\mu_{rc} = \exp(\gamma_r + x_{rc}^{\mathrm{T}}\beta)$:
  - the row parameters $\gamma_1, \ldots \gamma_R$ are **nuisance parameters**, not of interest;
  - we want inference for the **parameter of interest**, β.
- ☐ Corresponding multinomial model has fixed row totals m_r and probabilities

$$\pi_{rc} = \frac{\mu_{rc}}{\sum_{d=1}^{C} \mu_{rd}} = \frac{\exp(\gamma_r + x_{rc}^{\mathrm{T}}\beta)}{\sum_{d=1}^{C} \exp(\gamma_r + x_{rd}^{\mathrm{T}}\beta)} = \frac{\exp(x_{rc}^{\mathrm{T}}\beta)}{\sum_{d=1}^{C} \exp(x_{rd}^{\mathrm{T}}\beta)},$$

  for $r = 1, \ldots, R$, $c = 1, \ldots, C$; i.e., one multinomial variable for each row.
- ☐ The resulting multinomial log likelihood is

$$\begin{aligned} \ell_{\mathrm{Mult}}(\beta; y \mid m) &\equiv \sum_{r=1}^{R} \sum_{c=1}^{C} y_{rc} \log \pi_{rc} \\ &= \sum_{r=1}^{R} \left\{ \sum_{c=1}^{C} y_{rc} x_{rc}^{\mathrm{T}}\beta - m_r \log\left(\sum_{d=1}^{C} e^{x_{rd}^{\mathrm{T}}\beta} \right) \right\}. \end{aligned}$$

Contingency tables and Poisson response models, II

Lemma 41 *Show that if parameters τ_r for the row margins are included in the above setup, then we can write*

$$\ell_{\mathrm{Poiss}}(\beta, \tau) = \ell_{\mathrm{Poiss}}(\tau; m) + \ell_{\mathrm{Mult}}(\beta; y \mid m).$$

- ☐ Implications:
  - the MLEs of β and τ based on the LHS are the same as those from separate maximisations of the terms on the right:
    - ▷ $\widehat{\beta}$ equals the MLE for the multinomial model,
    - ▷ $\widehat{\tau}_r = m_r$
  - the observed and expected information matrices for β, τ are block diagonal.
  - SEs based on the multinomial and Poisson models are equal (exercise).
- ☐ General conclusion: inferences on β are the same for multinomial and Poisson models,

  provided the parameters associated to the margins fixed under the multinomial model, i.e., the γ_r, are included in the Poisson fit.

Note to Lemma 41

☐ The Poisson model has no conditioning, so the log likelihood is

$$\ell_{\rm Poiss}(\beta,\gamma) \equiv \sum_{r,c} (y_{rc}\log\mu_{rc} - \mu_{rc}) = \sum_{r=1}^R \left(m_r\gamma_r + \sum_{c=1}^C y_{rc}x_{rc}^{\rm T}\beta - e^{\gamma_r}\sum_{c=1}^C e^{x_{rc}^{\rm T}\beta} \right),$$

where we use the fact that $\log\mu_{rc} = \gamma_r + x_{rc}^{\rm T}\beta$.

☐ Now we reparametrise in terms of the row totals $\tau_r = \sum_c \mu_{rc}$, noting that

$$\tau_r = e^{\gamma_r}\sum_{c=1}^C e^{x_{rc}^{\rm T}\beta}, \quad \gamma_r = \log\tau_r - \log\left\{\sum_{c=1}^C \exp(x_{rc}^{\rm T}\beta)\right\},$$

so

$$\begin{aligned} \ell_{\rm Poiss}(\beta,\tau) &\equiv \sum_{r=1}^R (m_r\log\tau_r - \tau_r) + \sum_{r=1}^R \left\{ \sum_{c=1}^C y_{rc}x_{rc}^{\rm T}\beta - m_r\log\left(\sum_{c=1}^C e^{x_{rc}^{\rm T}\beta}\right)\right\}, \\ &= \ell_{\rm Poiss}(\tau; m) + \ell_{\rm Mult}(\beta; y \mid m), \end{aligned}$$

which is the log likelihood corresponding to

– independent Poisson row totals m_r with means τ_r, and, independent of this,

– the multinomial log likelihood for the contingency table.

Jacamar data

Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

	Aphrissa boisduvalli N/S/E	*Phoebis argante* N/S/E	*Dryas iulia* N/S/E	*Pierella luna* N/S/E	*Consul fabius* N/S/E	*Siproeta stelenes*† N/S/E
Unpainted	0/0/14	6/1/0	1/0/2	4/1/5	0/0/0	0/0/1
Brown	7/1/2	2/1/0	1/0/1	2/2/4	0/0/3	0/0/1
Yellow	7/2/1	4/0/2	5/0/1	2/0/5	0/0/1	0/0/3
Blue	6/0/0	0/0/0	0/0/1	4/0/3	0/0/1	0/1/1
Green	3/0/1	1/1/0	5/0/0	6/0/2	0/0/1	0/0/3
Red	4/0/0	0/0/0	6/0/0	4/0/2	0/0/1	3/0/1
Orange	4/2/0	6/0/0	4/1/1	7/0/1	0/0/2	1/1/1
Black	4/0/0	0/0/0	1/0/1	4/2/2	7/1/0	0/1/0

† includes *Philaethria dido* also.

Jacamar data: Models

- Let factors F, S, C represent the 3 fates, the 6 species, and the 8 colours.
- The models $C*S$, $C*S+F$, and $C*S+C*F$ mean we set

$$\log \mu_{csf} = \alpha_{cs}, \quad \log \mu_{csf} = \alpha_{cs} + \gamma_f, \quad \log \mu_{csf} = \alpha_{cs} + \gamma_{cf}.$$

- The vector of probabilities corresponding to the model with terms $C*S$ is

$$(\pi_{cs1}, \pi_{cs2}, \pi_{cs3}) = \left(\frac{\mu_{cs1}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs2}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs3}}{\sum_{f=1}^{3} \mu_{csf}} \right) = (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}),$$

  and that corresponding to the model with terms $C*S+F$ is

$$\begin{aligned} (\pi_{cs1}, \pi_{cs2}, \pi_{cs3}) &= \left(\frac{\mu_{cs1}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs2}}{\sum_{f=1}^{3} \mu_{csf}}, \frac{\mu_{cs3}}{\sum_{f=1}^{3} \mu_{csf}} \right) \\ &= \frac{1}{e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3}} \left(e^{\gamma_1}, e^{\gamma_2}, e^{\gamma_3} \right). \end{aligned}$$

- Exercise: similar computations for $C*S+C*F$ and $C*S+C*F+S*F$.

Jacamar data: Analysis of deviance

Deviances for log-linear models fitted to jacamar data.

Terms	df	Deviance
$C*S$	88	259.42
$C*S+F$	86	173.86
$C*S+C*F$	72	139.62
$C*S+S*F$	76	148.23
$C*S+C*F+S*F$	62	90.66
$C*S*F$	0	0

- The null model $C*S$ is not of interest.
- The first model it is sensible to fit is $C*S+F$.
- The best model seems to be $C*S+C*F+S*F$, corresponding to independent effects of species and colour, though its deviance is high (but remember the two outlying cells!)

Pneumoconiosis data

Period of exposure x and prevalence of pneumoconiosis amongst coalminers.

	Period of exposure (years)							
	5.8	15	21.5	27.5	33.5	39.5	46	51.5
Normal	98	51	34	35	32	23	12	4
Present	0	2	6	5	10	7	6	2
Severe	0	1	3	8	9	8	10	5

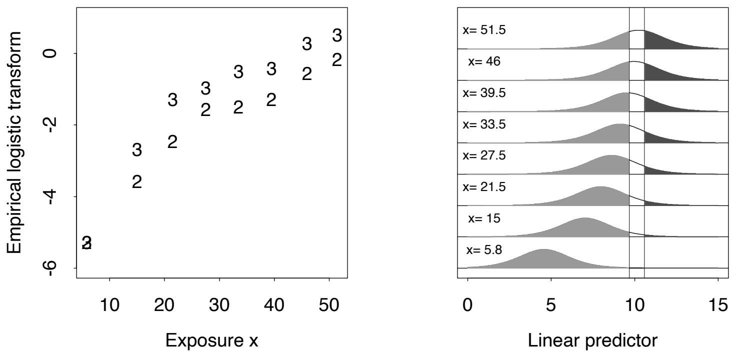
- Here
$$\text{Normal} \quad < \quad \text{Present} \quad < \quad \text{Severe},$$
so these are ordinal responses with $d = 3$ categories and the total in each group (corresponding to each period of exposure) fixed.
- It probably is reasonable to imagine that the choice of category stems from an underlying continuous variable, even if this cannot be quantified very well.

Models

- Assume we have n independent individuals whose responses $I_1, \dots, I_n$ fall into the set $\{1, \dots, d\}$, corresponding to d ordered categories, and that
$$\gamma_l = \mathrm{P}(I_j \le l) = \pi_1 + \cdots + \pi_l, \quad l = 1, \dots, d, \quad \gamma_d = 1,$$
- The corresponding likelihood is $\prod_{j=1}^n \pi_{I_j}$, where usually the contribution $\pi_{I_j} \equiv \pi_{I_j}(\eta_j)$ for individual j will depend on covariates x_j through a linear predictor $\eta_j = x_j^{\mathrm{T}}\beta$.
- We often want the interpretation of the parameters not to change if we merge adjacent categories, and we can do this using an underlying tolerance distribution, with
$$I_j = l \quad \Leftrightarrow \quad x_j^{\mathrm{T}}\beta + \varepsilon_j \in (\zeta_{l-1}, \zeta_l], \quad \zeta_0 = -\infty < \zeta_1 < \cdots < \zeta_{d-1} < \zeta_d = \infty,$$
where the tolerance distribution F of ε_j is often taken to be logistic, giving the **proportional odds model**, and
$$\pi_l(x_j^{\mathrm{T}}\beta) = \mathrm{P}(\zeta_{l-1} < x_j^{\mathrm{T}}\beta + \varepsilon \le \zeta_l) = F(\zeta_l - x_j^{\mathrm{T}}\beta) - F(\zeta_{l-1} - x_j^{\mathrm{T}}\beta), \quad l = 1, \dots, d;$$
here $\zeta_1, \dots, \zeta_{d-1}$ are aliased with an intercept β_0 and are not usually of interest.
- Another standard choice is $F(u) = 1 - \exp\{-\exp(u)\}$.
- To fit, we just apply IWLS to the multinomial likelihood $\prod_{j=1}^n \pi_{I_j}$.

Pneumoconiosis data

Pneumoconiosis data analysis, showing how the implied fitted logistic distributions depend on x. Left: plots of empirical logistic transforms for comparing categories 1 with $2+3$ and $1+2$ with 3; the nonlinearity suggests using $\log x$ as covariate. Right: fitted model, showing probabilities for the three groups with an underlying logistic distribution.



Comments on count data

☐ Log-linear models are mathematically elegant and useful defaults for count data, with close links to logistic regression, based on the relation between the Poisson and multinomial distributions.

☐ Interpretation of log-linear models can be difficult, especially for contingency tables, because marginal and conditional parameters cannot be disentangled.

☐ Other models exist that are less elegant mathematically, but have better interpretations in practice.

☐ Also possible to fit models for ordinal data, using multinomial models and tolerance distribution interpretation used for binomial data.

Overdispersion

- ☐ Often find that discrete response data are more variable than might be expected from a simple Poisson or binomial model, so we see
  - residual deviances larger than expected
  - residuals more variable than expected under the model

  but otherwise no evidence of systematic lack of fit
- ☐ This is **overdispersion**, perhaps due to effect of unmeasured explanatory variables on the responses.

UK monthly AIDS reports 1983–1992

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end of 1992
Year	Quarter	0†	1	2	3	4	5	6	···	≥14	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1988	1	31	80	16	9	3	2	8	···	6	174
	2	26	99	27	9	8	11	3	···	3	211
	3	31	95	35	13	18	4	6	···	3	224
	4	36	77	20	26	11	3	8	···	2	205
1989	1	32	92	32	10	12	19	12	···	2	224
	2	15	92	14	27	22	21	12	···	1	219
	3	34	104	29	31	18	8	6	···		253
	4	38	101	34	18	9	15	6	···		233
1990	1	31	124	47	24	11	15	8	···		281
	2	32	132	36	10	9	7	6	···		245
	3	49	107	51	17	15	8	9	···		260
	4	44	153	41	16	11	6	5	···		285
1991	1	41	137	29	33	7	11	6	···		271
	2	56	124	39	14	12	7	10	···		263
	3	53	175	35	17	13	11	2			306
	4	63	135	24	23	12	1				258
1992	1	71	161	48	25	5					310
	2	95	178	39	6						318
	3	76	181	16							273
	4	67	66								133

AIDS data

- ☐ UK monthly reports of AIDS diagnoses 1983–1992, with reporting delay up to several years!
- ☐ Example of incomplete contingency table (very common in insurance)
- ☐ Simple (chain-ladder) model: number of reports in row j and column k is Poisson, with mean

$$\mu_{jk} = \exp(\alpha_j + \beta_k).$$

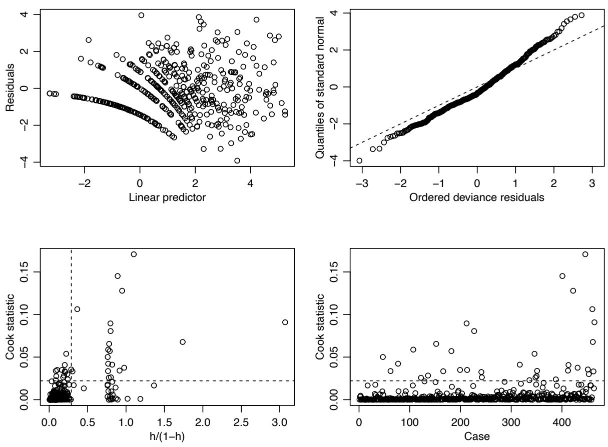
- ☐ Analysis of Deviance:

Model	df	Deviance reduction	df	Deviance
			464	14184.3
Time (rows)	37	6114.8	427	8069.5
Delay (cols)	14	7353.0	413	716.5

- ☐ Residual deviance is obviously far too large for a Poisson model to be OK, but the model is also too complex, since we expect smooth variation in the α_j.
- ☐ Next page shows residual analysis: no obvious problems, just generic overdispersion.

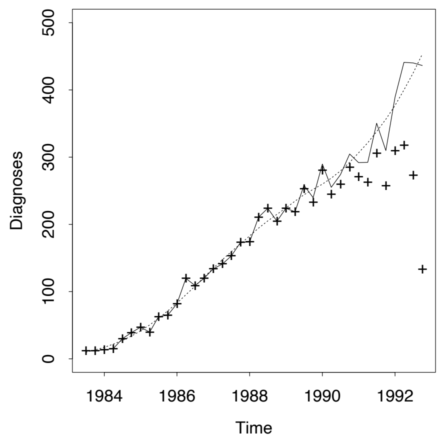
AIDS data: Assessment of fit

Diagnostic plots for fitted model: residuals against $\widehat{\eta}$ (top left); normal QQ-plot of residuals (top right); Cook statistic C_j against leverage ratio $h_j/(1-h_j)$ (lower left); Cook statistic C_j against case number (lower right).



AIDS data

- ☐ Data (+) and predicted true numbers based on simple Poisson model (solid) and GAM (dots).
- ☐ The Poisson model and data agree up to where data start to be missing.



Dealing with overdispersion

- ☐ Two basic approaches:
  - parametric modelling
  - quasi-likelihood estimation, based only on the variance function

Example 42 (Linear and quadratic variance functions) *Suppose that, conditional on* $\varepsilon > 0$, $Y \sim \text{Pois}(\mu\varepsilon)$, *where* $\text{E}(\varepsilon) = 1$ *and* $\text{var}(\varepsilon) = \xi$. *Show that this can lead to either linear or quadratic variance functions, but a lot of data may be needed to distinguish them.*

Comparison of variance functions for overdispersed count data. The linear and quadratic variance functions are $V_L(\mu) = (1 + \xi_L)\mu$ and $V_Q(\mu) = \mu(1 + \xi_Q\mu)$, with $\xi_L = 0.5$ and ξ_Q chosen so that $V_L(15) = V_Q(15)$.

μ	1	2	5	10	15	20	30	40	60
Linear	1.5	3.0	7.5	15.0	22.5	30	45	60	90
Quadratic	1.0	2.1	5.8	13.3	22.5	33	60	93	180

Note to Example 42

Let ε have unit mean and variance $\xi > 0$, and to be concrete suppose that conditional on ε, Y has the Poisson distribution with mean $\mu\varepsilon$. Then

$$\mathrm{E}(Y) = \mathrm{E}_\varepsilon \left\{ \mathrm{E}(Y \mid \varepsilon) \right\}, \quad \mathrm{var}(Y) = \mathrm{var}_\varepsilon \left\{ \mathrm{E}(Y \mid \varepsilon) \right\} + \mathrm{E}_\varepsilon \left\{ \mathrm{var}(Y \mid \varepsilon) \right\},$$

so the response has mean and variance

$$\mathrm{E}(Y) = \mathrm{E}_\varepsilon(\mu\varepsilon) = \mu, \quad \mathrm{var}(Y) = \mathrm{var}_\varepsilon(\mu\varepsilon) + \mathrm{E}_\varepsilon(\mu\varepsilon) = \mu(1 + \xi\mu).$$

If on the other hand the variance of ε is ξ/μ, then $\mathrm{var}(Y) = (1 + \xi)\mu$. In both cases the variance of Y is greater than its value under the standard Poisson model, for which $\xi = 0$. In the first case the variance function is quadratic, and in the second it is linear.

Negative binomial model

Example 43 (Negative binomial) *In Example 42, if ε is gamma with shape parameter $1/\nu$, show that*

$$f(y; \mu, \nu) = \frac{\Gamma(y + \nu)}{\Gamma(\nu) y!} \frac{\nu^\nu \mu^y}{(\nu + \mu)^{\nu + y}}, \quad y = 0, 1, \ldots, \quad \mu, \nu > 0,$$

and that quadratic and linear variance functions are obtained on setting $\nu = 1/\xi$ and $\nu = \mu/\xi$ respectively.
The log link function $\log \mu = x^{\mathrm{T}}\beta$ is most natural.
ξ is estimated by maximum likelihood or through Pearson's statistic.

Example 44 (AIDS data)

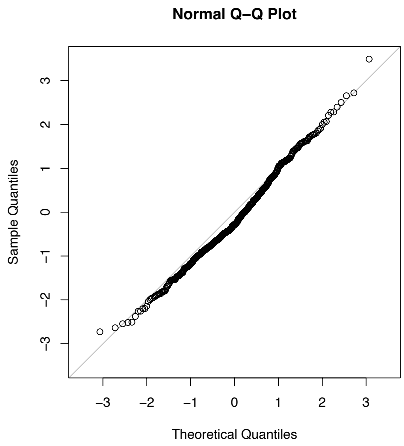
- ☐ *MLE $\widehat{\xi}_Q = 22.7\ (5.5)$*
- ☐ *Analysis of Deviance (with $\widehat{\xi}_Q$ fixed):*

Model	*df*	*Deviance reduction*	*df*	*Deviance*
			464	*7998.3*
Time (rows)	*37*	*3582.5*	*427*	*4415.8*
Delay (cols)	*14*	*3892.2*	*413*	*523.6*

- ☐ *Still somewhat overdispersed?*

AIDS data: Deviance residuals for NB model

Clear improvement over previous plots, even if not perfect.



Quasi-likelihood

- Recall two basic assumptions for the linear model:
  - the responses are uncorrelated with means $\mu_j = x_j^{\mathrm{T}}\beta$ and equal variances σ^2;
  - in addition to this, the responses are normally distributed.
- To avoid parametric modelling, we generalise the second-order assumptions, to

$$\mathrm{E}(Y_j) = \mu_j, \quad \mathrm{var}(Y_j) = \phi_j V(\mu_j), \quad g(\mu_j) = \eta_j = x_j^{\mathrm{T}}\beta,$$

  where the variance function $V(\cdot)$ and the link function are taken as known.
- We obtain estimates $\tilde{\beta}$ by solving the estimating equation

$$h(\beta; Y) = X^{\mathrm{T}} u(\beta) = \sum_{j=1}^{n} x_j u_j(\beta) = \sum_{j=1}^{n} x_j \frac{Y_j - \mu_j}{g'(\mu_j)\phi_j V(\mu_j)} = 0.$$

- If the mean structure is correct, then $\mathrm{E}(Y_j) = \mu_j$, so $\mathrm{E}\{h(\beta; Y)\} = 0$, and under mild conditions $\tilde{\beta}$ is consistent (but maybe not efficient) as $n \to \infty$.

Quasi-likelihood II

Recall that the general variance of an estimator $\tilde\beta$ defined by an estimating equation $h(\beta;Y)=0$ has sandwich form

$$\mathrm{E}\left\{-\frac{\partial h(\beta;Y)}{\partial\beta^{\mathrm{T}}}\right\}^{-1}\mathrm{var}\left\{h(\beta;Y)\right\}\mathrm{E}\left\{-\frac{\partial h(\beta;Y)^{\mathrm{T}}}{\partial\beta}\right\}^{-1}.$$

Lemma 45 *If $V(\mu)$ is correctly specified, then $\mathrm{var}\{\tilde\beta\}\doteq(X^{\mathrm{T}}WX)^{-1}$, where W is diagonal with (j,j) element $\{g'(\mu_j)^2\phi_jV(\mu_j)\}^{-1}$.*

☐ If $\phi_j=\phi a_j$, with known $a_j>0$ and unknown $\phi>0$, then we obtain
  - $\tilde\beta$ by fitting the GLM with variance function $V(\mu)$ and link $g(\mu)$;
  - standard errors by multiplying the standard errors for this fit by $\widehat{\phi}^{1/2}$, where

$$\widehat{\phi}=\frac{1}{n-p}\sum_{j=1}^{n}\frac{(y_j-\widehat{\mu}_j)^2}{a_jg'(\mu_j)^2V(\widehat{\mu}_j)}.$$

Note to Lemma 45

- ☐ Note first that we can write

$$u_j(\beta) \equiv u_j(\mu_j) = \frac{A_j(\mu_j)}{B_j(\mu_j)},$$

where $A_j(\mu_j) = Y_j - \mu_j$ and $B_j(\mu_j) = g'(\mu_j)\phi_j V(\mu_j)$ and only A_j is random, and $\mathrm{E}\{A_j(\mu_j)\} = 0$. Hence if we let prime denote derivative with respect to μ_j,

$$\frac{\partial u_j(\mu_j)}{\partial \mu_j} = \frac{A_j'(\mu_j)}{B_j(\mu_j)} - \frac{A_j(\mu_j)B_j'(\mu_j)}{B_j^2(\mu_j)}$$

has expectation $\mathrm{E}\{A_j'(\mu_j)\}/B_j(\mu_j) = -1/B_j(\mu_j)$.

- ☐ We require $\mathrm{E}\{-\partial h(\beta;Y)/\partial\beta^{\mathrm{T}}\}$ and $\mathrm{var}\{h(\beta;Y)\}$. Now

$$\frac{\partial u_j(\beta)}{\partial \beta^{\mathrm{T}}} = \frac{\partial \eta_j}{\partial \beta^{\mathrm{T}}}\frac{\partial \mu_j}{\partial \eta_j}\frac{\partial u_j(\beta)}{\partial \mu_j} = x_j^{\mathrm{T}}\frac{1}{g'(\mu_j)}u_j'(\mu_j),$$

which gives

$$\mathrm{E}\left\{-\frac{\partial h(\beta;Y)}{\partial \beta^{\mathrm{T}}}\right\} = -\sum_{j=1}^{n} x_j \mathrm{E}\left\{\frac{\partial u_j(\beta)}{\partial \beta^{\mathrm{T}}}\right\} = \sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \frac{1}{g'(\mu_j)^2\phi_j V(\mu_j)} = X^{\mathrm{T}}WX,$$

where W is the $n \times n$ diagonal matrix with $j\mathrm{th}$ element $\{g'(\mu_j)^2\phi_j V(\mu_j)\}^{-1}$. Moreover if in addition the variance function has been correctly specified, then $\mathrm{var}(Y_j) = \phi_j V(\mu_j)$, and hence

$$\mathrm{var}\{h(\beta;Y)\} = X^{\mathrm{T}}\mathrm{var}\{u(\beta)\}X = \sum_{j=1}^{n} x_j x_j^{\mathrm{T}} \frac{\mathrm{var}(Y_j)}{g'(\mu_j)^2\phi_j^2 V(\mu_j)^2} = X^{\mathrm{T}}WX.$$

Thus the sandwich equals $(X^{\mathrm{T}}WX)^{-1}$.

- ☐ Had the variance function been wrongly specified, the variance matrix of $\tilde{\beta}$ would have been of form $(X^{\mathrm{T}}WX)^{-1}(X^{\mathrm{T}}W'X)(X^{\mathrm{T}}WX)^{-1}$, where W' is a diagonal matrix involving the true and assumed variance functions. Only if the variance function has been chosen very badly will this sandwich matrix differ greatly from $(X^{\mathrm{T}}WX)^{-1}$, which therefore provides useful standard errors unless a plot of absolute residuals against fitted means is markedly non-random. In that case the choice of variance function should be reconsidered.

Quasi-likelihood III

- Under an exponential family model, $h(\beta;Y)$ is the score statistic, so $\tilde{\beta}$ is the MLE and is efficient (i.e., it has the smallest possible variance in large samples).
- If not, inference is valid provided g and V are correctly chosen, and $\tilde{\beta}$ is optimal among estimators based on linear combinations of the $Y_j - \mu_j$, by extending the Gauss–Markov theorem.
- In fact we can define a **quasi-likelihood** Q and its score through

$$Q(\beta;Y) = \sum_{j=1}^{n} \int_{Y_j}^{\mu_j} \frac{Y_j - u}{\phi a_j V(u)}\,\mathrm{d}u, \quad h(\beta;Y) = \frac{\partial}{\partial \beta} Q(\beta;Y),$$

  and a (quasi-)deviance as $D = -2\phi Q(\beta;Y)$.
- To compare models A, B with numbers of parameters $p_B < p_A$ and deviances $D_B > D_A$, we use the fact that

$$\frac{(D_B - D_A)/(p_A - p_B)}{\widehat{\phi}_A} \quad \dot{\sim} \quad F_{p_A - p_B, n-p_A},$$

  if the simpler model B is adequate. This is easy in R.

AIDS example

```
> aids.ql <- glm(y~factor(time)+factor(delay),family=quasipoisson,data=aids.in)
> anova(aids.ql,test="F")
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: y

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                            464    14184.3
factor(time)  37   6114.8       427     8069.5  92.638 < 2.2e-16 ***
factor(delay) 14   7353.0       413      716.5 294.402 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

- ☐ Overdispersion is widespread in count and proportion data.
- ☐ We deal with it either by
  - parametric modelling, or
  - quasi-likelihood (QL) estimation, which involves assumptions only on the mean-variance relationship.
- ☐ QL estimators equal the ML ones, but SEs are inflated by $\widehat{\phi}^{1/2}$.
- ☐ (Quasi-)deviance can also be defined, and used for model comparison, with F tests replacing χ^2 tests.

Generalized additive model

- ☐ Now we write

$$\mathrm{E}(y) = \mu, \quad g(\mu) = \eta = B\gamma = X\beta + Zb,$$

  where
  - y follows a GLM (or more general) distribution,
  - $g(\cdot)$ is a link function,
  - the rest is as before ...

  giving a **generalized additive model (GAM)**.
- ☐ For a general treatment, suppose we have a penalized log likelihood,

$$\ell_\lambda(\gamma) = \ell(\gamma) - \tfrac{1}{2}\gamma^{\mathrm{T}} D_\lambda \gamma = \sum_{j=1}^{n} \ell_j\{\eta_j(\gamma)\} - \tfrac{1}{2}\gamma^{\mathrm{T}} D_\lambda \gamma,$$

  where $\gamma_{m\times 1}$ (with $m = p + q$) contains both 'fixed-effects' $\beta_{p\times 1}$ and 'random-effects' $b_{q\times 1}$, the latter penalized using a symmetric positive semidefinite $m \times m$ matrix D_λ, and the underlying observations $y_1, \ldots, y_n$ giving likelihood contributions $\ell_1, \ldots, \ell_n$ are assumed to be independent.
- ☐ Now we apply the argument leading to the IWLS algorithm to ℓ_λ, leading to the **penalized iterative weighted least squares (PIWLS)** algorithm.

PIWLS

- ☐ For fixed λ, we apply (ridge regression) iterative weighted least squares with update step

$$\widehat{\gamma}_\lambda = (B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}} W z,$$

  where D_λ is the penalty matrix, and

$$\begin{aligned}
B_{n\times m} &= \partial\eta/\partial\gamma^{\mathrm{T}}, \quad \text{(design matrix)}\\
W_{n\times n} &= \mathrm{diag}(w_1, \ldots, w_n), \quad w_j = \{\mathrm{E}(-\partial^2 \ell_j/\partial\eta_j^2)\}, \quad \text{(weights)}\\
u_{n\times 1} &= \partial\ell/\partial\eta, \quad \text{(score vector)},\\
z_{n\times 1} &= B\gamma + W^{-1}u, \quad \text{(adjusted dependent variable)}.
\end{aligned}$$

  It is easier (but less stable) to use the (random) $-\partial^2\ell_j/\partial\eta_j^2$ in place of $\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)$.
- ☐ Thus to obtain (penalized) MLEs $\widehat{\gamma}_\lambda$ we use the **PIWLS algorithm**:
- ☐ fix λ and take an initial $\widehat{\gamma}_\lambda$. Repeat
  - compute η, B, W, u, z;
  - compute new $\widehat{\gamma}_\lambda = (B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}} W z$;

  until changes in $\ell_\lambda(\widehat{\gamma}_\lambda)$ (or $\widehat{\gamma}_\lambda$, or both) are lower than some tolerance.
- ☐ We may add a line search: if $\ell_\lambda(\widehat{\gamma}_{\lambda,\mathrm{new}}) < \ell_\lambda(\widehat{\gamma}_{\lambda,\mathrm{old}})$, halve the step length and try again.

Note: Derivation of PIWLS algorithm

☐ To find the estimate $\widehat{\gamma}_\lambda$ starting from a trial value γ, we make a Taylor series expansion in the score equation

$$0 = \frac{\partial \ell_\lambda(\widehat{\gamma}_\lambda)}{\partial \gamma} \doteq \frac{\partial \ell_\lambda(\gamma)}{\partial \gamma} + \frac{\partial^2 \ell_\lambda(\gamma)}{\partial \gamma \partial \gamma^{\mathrm{T}}}(\widehat{\gamma}_\lambda - \gamma),$$

where

$$\frac{\partial \ell_\lambda(\gamma)}{\partial \gamma} = B^{\mathrm{T}} u(\gamma) - D_\lambda \gamma, \quad \frac{\partial^2 \ell_\lambda(\gamma)}{\partial \gamma_r \partial \gamma_s} = \sum_{j=1}^n \frac{\partial \eta_j(\gamma)}{\partial \gamma_r} \frac{\partial^2 \ell_j(\gamma)}{\partial \eta_j^2} \frac{\partial \eta_j(\gamma)}{\partial \gamma_s} + \sum_{j=1}^n \frac{\partial^2 \eta_j(\gamma)}{\partial \gamma_r \partial \gamma_s} u_j(\gamma) + D_{\lambda,r,s},$$

where $B \equiv B(\gamma) = \partial \eta / \partial \gamma^{\mathrm{T}}$. If we use the approximation

$$-\frac{\partial^2 \ell_\lambda(\gamma)}{\partial \gamma \partial \gamma^{\mathrm{T}}} \doteq B^{\mathrm{T}} W B + D_\lambda, \quad W = \mathrm{diag}\left\{-\mathrm{E}\left(\partial^2 \ell_j / \partial \eta_j^2\right)\right\},$$

where the matrix is replaced by its expectation, then

$$\begin{aligned} 0 &\doteq B^{\mathrm{T}} u(\gamma) - D_\lambda \gamma - (B^{\mathrm{T}} W B + D_\lambda)(\widehat{\gamma}_\lambda - \gamma) \\ &= B^{\mathrm{T}} u(\gamma) + B^{\mathrm{T}} W B \gamma - (B^{\mathrm{T}} W B + D_\lambda)\widehat{\gamma}_\lambda. \end{aligned}$$

If $B^{\mathrm{T}} W B + D_\lambda$ is invertible, this gives

$$\widehat{\gamma}_\lambda \doteq (B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}}(u + W B \gamma) = (B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}} W z,$$

where $z = B\gamma + W^{-1} u$, as required.

Relation with least squares

☐ With fixed λ, the penalized MLE

$$\widehat{\gamma}_\lambda = (B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}} W z$$

results from fixing γ, and then iteratively solving the minimization problem

$$\min_\gamma \left\| \begin{pmatrix} W^{1/2} z \\ 0 \end{pmatrix}_{(n+m)\times 1} - \begin{pmatrix} W^{1/2} B \\ Q_\lambda \end{pmatrix}_{(n+m)\times m} \gamma_{m\times 1} \right\|^2,$$

where Q_λ is a matrix square root of D_λ, i.e., $Q_\lambda^{\mathrm{T}} Q_\lambda = D_\lambda$.

☐ The corresponding smoothing matrix is taken to be

$$S_\lambda = B(B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}} W,$$

and the effective degrees of freedom for a smooth component are taken as the sum of the relevant diagonal elements of the matrix

$$A = (B^{\mathrm{T}} W B + D_\lambda)^{-1} B^{\mathrm{T}} W B,$$

with both S_λ and A evaluated at the final step of the iteration.

Numerical example from Wood (2011, JRSSB)

The usual methods (AIC, GCV, ...) for choosing λ are available, but we focus on likelihood methods; see below.

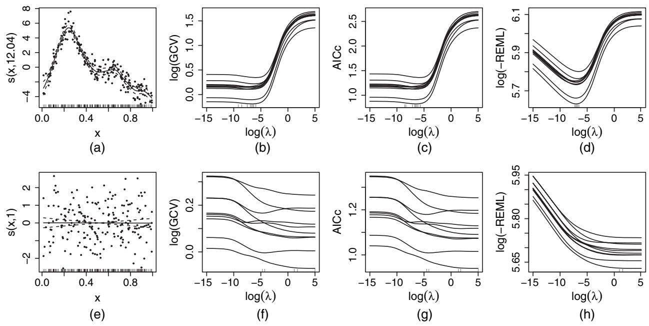


Fig. 1. Example comparison of GCV, AICc and REML criteria: (a) some (x,y)-data modelled as $y_i = f(x_i) + \varepsilon_i$, ε_i independent and identically distributed $N(0,\sigma^2)$ where smooth function f was represented by using a rank 20 thin plate regression spline (Wood, 2003); (b)–(d) various smoothness selection criteria plotted against logarithmic smoothing parameters, for 10 replicates of the data (each generated from the same 'truth') (note how shallow the GCV and AICc minima are relative to the sampling variability, resulting in rather variable optimal λ-values (which are shown as a rug plot), and a propensity to undersmooth; in contrast the REML optima are much better defined, relative to the sampling variability, resulting in a smaller range of λ-estimates); (e)–(h) are equivalent to (a)–(d), but for data with no signal, so that the appropriate smoothing parameter should tend to ∞ (note GCV's and AICc's occasional multiple minima and undersmoothing in this case, compared with the excellent behaviour of REML; ML (which is not shown) has a similar shape to REML)

Approaches to iteration

- Once we have an iterative approach to estimating λ for fixed γ, there are two main algorithms:
  - **performance iteration** — repeat { fix λ, update γ with one step of PIWLS, update λ } to convergence;
  - **outer iteration** — repeat { fix λ, iterate PIWLS to convergence, update λ } to convergence.
- Performance iteration
  - is computationally efficient,
  - but since the objective function for γ changes at each step, it may not converge—especially in the context of **concurvity** (collinearity for curves ...), when two or more smooth functions are (almost) confounded.
- Outer iteration
  - is computationally more burdensome,
  - but will converge to a (local) optimum.

Choice of λ

- The choice of λ can be based on the marginal density of y,

$$f(y;\beta,\lambda) = \int f(y \mid b;\beta) f(b;\lambda)\,\mathrm{d}b,$$

which has no closed form in general (but is Gaussian if both fs are Gaussian).
- Various ways to approximate the integral:
  - quadrature (doesn't work well when $\dim(b)$ is high);
  - simulation (e.g., importance sampling, same problems as quadrature);
  - Laplace approximation;
  - use the EM algorithm to avoid approximating the integral.
- We focus on Laplace approximation.

Laplace approximation

Lemma 46 *Let $h(u)$ be a smooth convex function defined for $u \in \mathbb{R}$, with a minimum at $u = \tilde{u}$, where $h'(\tilde{u}) = 0$ and $h''(\tilde{u}) > 0$, and let*

$$I_n = \int_{-\infty}^{\infty} e^{-nh(u)}\,\mathrm{d}u.$$

Then if we write $h_2 = h''(\tilde{u})$, $h_3 = h'''(\tilde{u})$, etc., we have

$$I_n = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})} \times \left\{1 + n^{-1}\left(\frac{5h_3^2}{24h_2^3} - \frac{h_4}{8h_2^2}\right) + O\left(n^{-2}\right)\right\}.$$

The **Laplace approximation** to the integral is

$$\tilde{I}_n = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})}.$$

Note on Lemma 46

- ☐ Close to $\tilde{u}$ a Taylor series expansion gives

$$h(u) \doteq h(\tilde{u}) + h'(\tilde{u})(u-\tilde{u}) + \tfrac{1}{2}h''(\tilde{u})(u-\tilde{u})^2 = h(\tilde{u}) + \tfrac{1}{2}h_2(u-\tilde{u})^2$$

so if we set $z = (nh_2)^{1/2}(u-\tilde{u})$ then $u = \tilde{u} + (nh_2)^{1/2}z$, $\mathrm{d}u/\mathrm{d}z = (nh_2)^{-1/2}$, and arguing heuristically (ignoring the third and higher terms),

$$\begin{aligned} I_n &\doteq e^{-nh(\tilde{u})}\int_{-\infty}^{\infty} e^{-nh_2(u-\tilde{u})^2/2}\,\mathrm{d}u \\ &= e^{-nh(\tilde{u})}\int_{-\infty}^{\infty} e^{-z^2/2}\frac{\mathrm{d}u}{\mathrm{d}z}\,\mathrm{d}z \\ &= \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})}, \end{aligned}$$

because the normal density has unit integral.

- ☐ A more detailed accounting is needed to get the error term. We start by writing

$$\begin{aligned} nh(u) &\doteq nh(\tilde{u}) + \tfrac{1}{2}nh_2(u-\tilde{u})^2 + \tfrac{1}{6}nh_3(u-\tilde{u})^3 + \tfrac{1}{24}nh_4(u-\tilde{u})^4 + \cdots \\ &= nh(\tilde{u}) + \tfrac{1}{2}z^2 + \tfrac{1}{6}\frac{h_3/h_2^{3/2}}{n^{1/2}}z^3 + \tfrac{1}{24}\frac{h_4/h_2^2}{n}z^4 + O(n^{-3/2}) \\ &= nh(\tilde{u}) + \tfrac{1}{2}z^2 + \frac{A}{n^{1/2}}z^3 + \frac{B}{n}z^4 + O(n^{-3/2}) \end{aligned}$$

say. Hence

$$\begin{aligned} e^{-nh(u)} &= e^{-nh(\tilde{u})-\frac{1}{2}z^2}\left\{1 - \frac{A}{n^{1/2}}z^3 - \frac{B}{n}z^4 + \tfrac{1}{2}\left(-\frac{A}{n^{1/2}}z^3 - \frac{B}{n}z^4\right)^2 + O(n^{-3/2})\right\} \\ &= e^{-nh(\tilde{u})-\frac{1}{2}z^2}\left\{1 - \frac{A}{n^{1/2}}z^3 - \frac{B}{n}z^4 + \tfrac{1}{2}\frac{A^2}{n}z^6 + O(n^{-3/2})\right\}. \end{aligned}$$

- ☐ As the odd moments of the normal density are zero, integration with respect to z leaves only the n^{-1} term and the next remaining term is $O(n^{-2})$. The fourth and sixth moments of the standard normal distribution are respectively 3 and 15, and

$$15A^2/2 - 3B = 15(h_3/h_2^{3/2}/6)^2/2 - 3\{h_4/(24h_2)\} = \frac{15h_3^2}{72h_2^3} - \frac{h_4}{8h_2^2} = \frac{5h_3^2}{24h_2^3} - \frac{h_4}{8h_2^2},$$

as required.

Comments

- ☐ The $O(1/n)$ error is relative, so the approximation is often surprisingly accurate;
- ☐ since the odd moments of the normal density are all zero, the expansion has only terms whose orders are even powers of $n^{-1/2}$, i.e., $n^{-1}, n^{-2}, \ldots$;
- ☐ $\tilde{I}_n$ involves only h and the second derivative h'' at $\tilde{u}$, so is very easy to obtain, numerically if necessary;
- ☐ the series is asymptotic, so the partial sums may not converge, and including more terms than the leading term may not be useful;
- ☐ as most of the normal probability lies within ± 3 standard deviations of the mean, the limits of the integral are almost irrelevant provided they lie outside the interval $\tilde{u} \pm 3(nh_2)^{-1/2}$;
- ☐ if

$$I_n = \int_{-\infty}^{\infty} e^{-nh(u)}\,\mathrm{d}u, \quad J_n = \int_{-\infty}^{\infty} e^{-nh^*(u)}\,\mathrm{d}u,$$

where $h^*(u) = h(u) + O(n^{-1})$, then

$$(I_n/J_n) \div (\tilde{I}_n/\tilde{J}_n) = 1 + O(n^{-2}),$$

so two Laplace approximations can be better than one.

Laplace approximation, bis

Lemma 47 *Let $h(u)$ be a smooth convex function defined for $u \in \mathbb{R}^m$, with a minimum at $u = \tilde{u}$, where $\partial h(\tilde{u})/\partial u = 0$ and the matrix of partial derivatives $h_2 \equiv \partial^2 h(\tilde{u})/\partial u \partial u^{\mathrm{T}}$ is positive definite, and let*

$$I_n = \int_{\mathbb{R}^m} e^{-nh(u)}\,\mathrm{d}u.$$

Then

$$I_n = \tilde{I}_n\left\{1 + O(n^{-1})\right\} = \left(\frac{2\pi}{n}\right)^{m/2} |h_2|^{-1/2} e^{-nh(\tilde{u})}\left\{1 + O(n^{-1})\right\}.$$

- ☐ To apply this, we impose $\gamma = (\beta_{p\times 1}^{\mathrm{T}}, b_{q\times 1}^{\mathrm{T}})^{\mathrm{T}} \sim \mathcal{N}_m(0, D_\lambda^{-1})$, and write

$$f(y;\lambda) = \int f(y;\gamma) f(\gamma;\lambda)\,\mathrm{d}\gamma = \frac{|D_\lambda|_+^{1/2}}{(2\pi)^{m/2}} \int \exp\left\{\ell_\lambda(\gamma)\right\}\,\mathrm{d}\gamma,$$

where β is unpenalised, $|D_\lambda|_+$ is the product of the non-negative eigenvalues of D_λ, and

$$\ell_\lambda(\gamma) = \ell(\gamma) - \tfrac{1}{2}\gamma^{\mathrm{T}} D_\lambda \gamma;$$

the assumptions of Lemma 47 should be satisfied by $h(u) \equiv -n^{-1}\ell_\lambda(\gamma)$.

Approximate REML

- Laplace approximation gives the approximate restricted log likelihood

$$\ell_{\rm p}(\lambda) \equiv \tfrac{1}{2}\log|D_\lambda|_+ - \tfrac{1}{2}\log|B^{\rm T}WB^{\rm T} + D_\lambda| + \ell(\widehat\gamma_\lambda) - \tfrac{1}{2}\widehat\gamma_\lambda^{\rm T} D_\lambda \widehat\gamma_\lambda + O_p(n^{-1}),$$

where $O_p(n^{-1})$ is a (random) term of order n^{-1} and

$$\widehat\gamma_\lambda = (B^{\rm T}WB + D_\lambda)^{-1}B^{\rm T}Wz$$

results from iterating PIWLS to convergence for fixed λ and satisfies $\partial\ell_\lambda(\widehat\gamma_\lambda)/\partial\gamma = 0$.
- The expression for $\widehat\gamma_\lambda$ contains

$$B \equiv B(\widehat\gamma_\lambda), \quad W \equiv W(\widehat\gamma_\lambda), \quad z = B(\widehat\gamma_\lambda)\widehat\gamma_\lambda + W^{-1}(\widehat\gamma_\lambda)u(\widehat\gamma_\lambda),$$

which involve the first two derivatives of the log likelihood contributions ℓ_j.
- Newton–Raphson maximization of $\ell_{\rm p}(\lambda)$ requires its first two derivatives, so we need

$$\frac{\partial\widehat\gamma_\lambda}{\partial\lambda}, \quad \frac{\partial^2\widehat\gamma_\lambda}{\partial\lambda\partial\lambda^{\rm T}},$$

which will involve the third and fourth derivatives of the ℓ_j ...could be painful.
- A version of this is implemented in `mgcv`.

Example

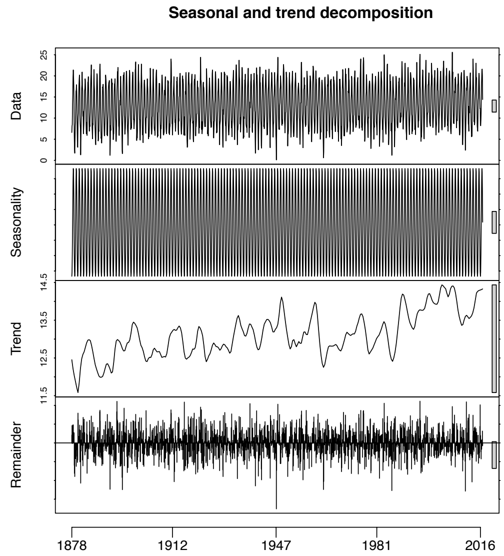
- Central England temperature time series, monthly maxima, 1878–2016, to which we fit the **generalized extreme-value (GEV) distribution**,

$$F(y;\mu,\sigma,\xi) = \exp\left[-\{1+\xi(y-\mu)/\sigma\}_+^{-1/\xi}\right], \quad y \in \mathbb{R}, \mu, \xi \in \mathbb{R}, \sigma > 0,$$

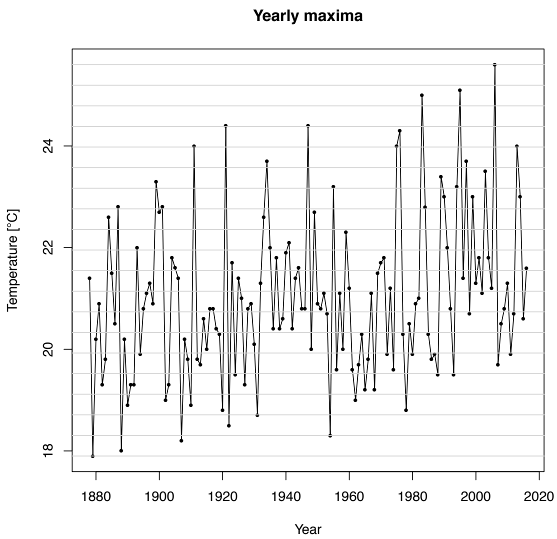
where $a_+ = \max(a, 0)$. Here
  - μ is a location parameter,
  - σ is a scale parameter,
  - ξ is a shape parameter, with $\xi = 0$ corresponding to the Gumbel distribution function $\exp[-\exp\{-(y-\mu)/\sigma\}]$, and $\xi < 0$ giving a distribution with upper maximum $\mu - \sigma/\xi$.
- The data show strong seasonality and some trend, but also large annual variation.
- We use the ideas from above to fit a model with
  - $\mu = s(\text{month}) + s(\text{year})$,
  - $\sigma = s(\text{month})$,
  - $\xi = s(\text{month})$,

  with results shown on the following slides.

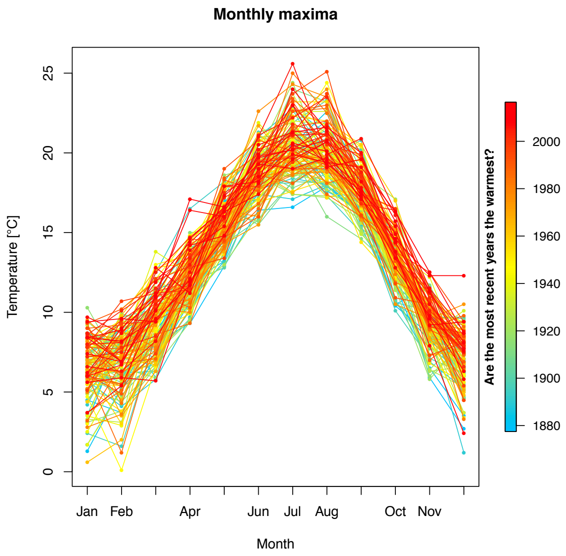
Example: STL decomposition



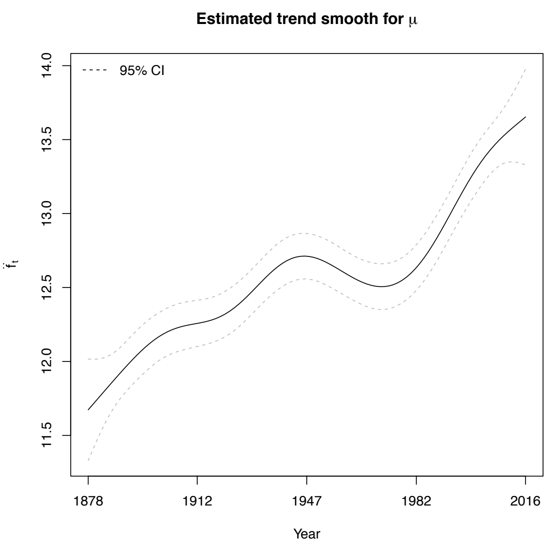
Example: Annual maxima



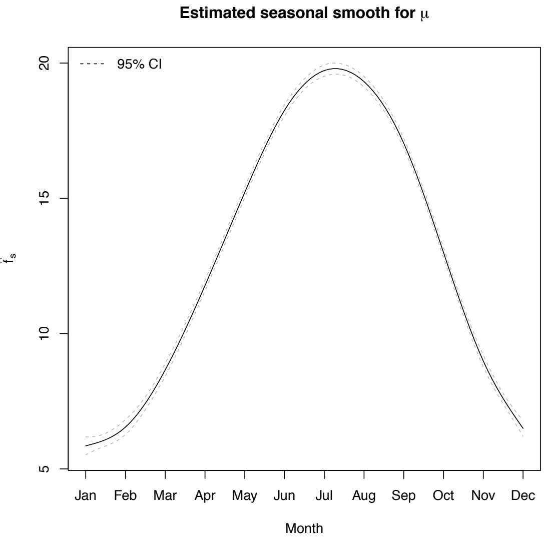
Example: Monthly maxima



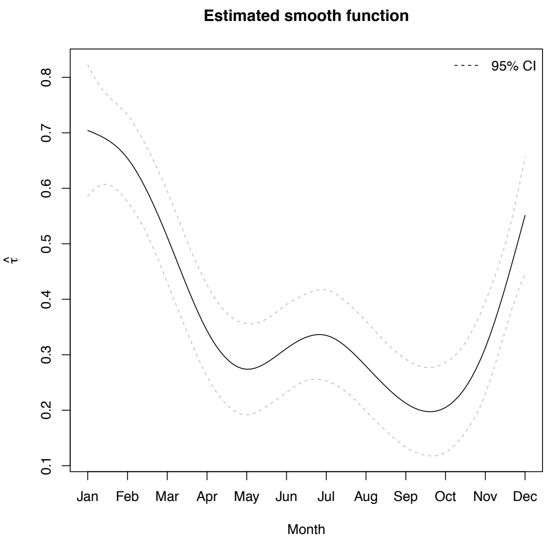
Example: Trend in μ



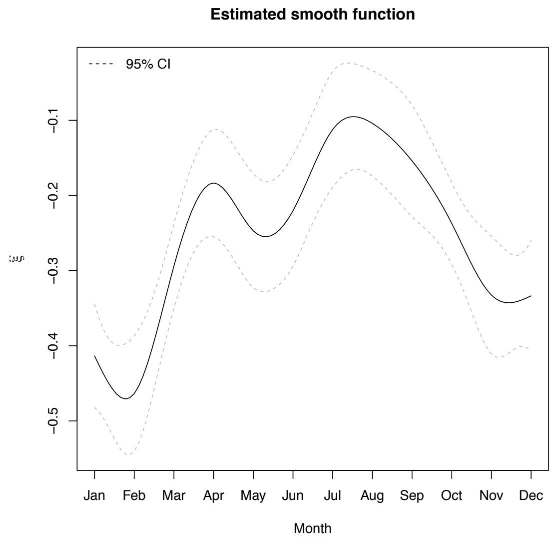
Example: Seasonality in μ



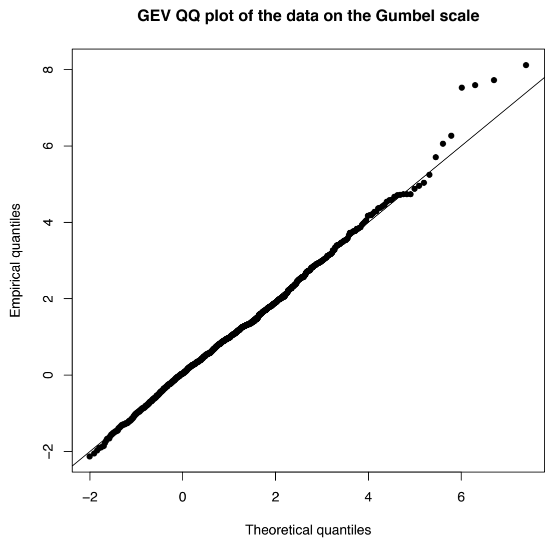
Example: Seasonality in σ



Example: Seasonality in ξ



Example: Residuals



Closing

- ☐ The basic ideas of regression, dependence of a response on covariables, extend far beyond the Gaussian linear model, to
  - general response distributions (Poisson, binomial, GEV, ...);
  - random effects models—some parameters treated as random, and others as fixed;
  - smooth curve fitting by basis function methods in (generalized) additive models.
- ☐ Unifying themes are:
  - (semi-)parametric modelling;
  - maximum likelihood estimation, performed by
  - variants of (iterative weighted) least squares algorithms;
  - residuals and other diagnostics;
  - penalized likelihood estimation in presence of random effects;
  - best linear unbiased prediction (BLUP).