Regression Methods

Anthony Davison

http://stat.epfl.ch

1 The Linear Model

slide 2

1.1 Introduction

slide 3

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 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, ...) 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 $\partial^2 Q/\partial\beta_r\partial\gamma_s$.
- ☐ $u \perp v$ means that the vectors u and v are orthogonal (i.e., $u^{\mathrm{T}}v = 0$); ditto for matrices.
- ☐ $Y \perp\!\!\!\perp Z$ means that the random variables Y and Z are independent.

Useful matrix decompositions

- ☐ **Singular value decomposition (SVD)**: any real matrix X can be written in the form

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

  where
  - $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_p)$ are orthogonal (i.e., $U^{\mathrm{T}}U = UU^{\mathrm{T}} = I_n$, $V^{\mathrm{T}}V = VV^{\mathrm{T}} = I_p$) and D is $n \times p$ rectangular diagonal with real 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.
- ☐ The SVD implies that the ranks of X, $X^{\mathrm{T}}X$ and XX^{T} are equal and at most m.
- ☐ **Spectral theorem**: any real symmetric matrix H can be written as

$$H_{n\times n} = U_{n\times n} D_{n\times n} U^{\mathrm{T}}_{n\times n},$$

  where
  - $D = \mathrm{diag}(d_1, \ldots, d_n)$ contains the eigenvalues of H;
  - U is an orthogonal matrix whose columns are the corresponding eigenvectors; and
  - if H is positive semi-definite then $d_1 \geq \cdots \geq d_n \geq 0$.

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).
- 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 *When X has rank p and $n \geq p$ 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 has rank p, so too does the $p \times p$ matrix $X^{\mathrm{T}}X$, which is therefore invertible.
- 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 $X^{\mathrm{T}}X$ is invertible 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 the span of the columns of X, $\mathrm{span}(X)$.
- The inner product between $\widehat{y}$ and $y - \widehat{y}$ equals zero, because $\widehat{y} = Hy$, $y - \widehat{y} = (I - H)y$, and $\widehat{y}^{\mathrm{T}}(y - \widehat{y}) = y^{\mathrm{T}}H^{\mathrm{T}}(I - H)y = y^{\mathrm{T}}(H - H)y = 0$. Hence $\widehat{y}$ and $y - \widehat{y}$ are orthogonal.
- Clearly $HX = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X = X$, so $H(X\beta) = X\beta$ for any $\beta \in \mathbb{R}^p$, i.e., a vector in $\mathrm{span}(X)$ is left unchanged by multiplication by H.

Analysis of variance I

Lemma 2 *Let $X_{n\times p} = (X_0, X_1, \ldots, X_R)$ have rank p, where $p \le n$, and let H_r denote the projection matrices formed using $(X_0, \ldots, X_r)$, for $r = 0, \ldots, R$; hence $H_R = H$. Define $P_r = H_r - H_{r-1}$ for $r = 1, \ldots, R$ and $P_{R+1} = I - H$. Then (i) $H_r H_s = H_r$ whenever $r \le s$, (ii) $H_0 P_r = 0$ for any r, and (iii) the matrices P_r are symmetric and idempotent, with $P_r P_s = 0$ when $r \ne s$.*

- In the setup of Lemma 2 suppose we fit the models with projection matrices $H_0, \ldots, H_R = H$ and corresponding fitted values $\widehat{y}_r = H_r y$. Then

$$\begin{aligned} y &= \widehat{y}_0 + (\widehat{y}_1 - \widehat{y}_0) + \cdots + (\widehat{y}_R - \widehat{y}_{R-1}) + (y - \widehat{y}_R) \\ &= H_0 y + (H_1 - H_0)y + \cdots + (H_R - H_{R-1})y + (I - H)y \\ &= H_0 y + P_1 y + \cdots + P_R y + P_{R+1} y, \end{aligned}$$

  and Lemma 2 implies that the terms on the RHS are orthogonal, i.e.,

$$(H_0 y)^{\mathrm{T}}(P_r y) = 0, \quad (P_s y)^{\mathrm{T}}(P_r y) = 0, \quad r \ne s.$$

- Hence Pythagoras' theorem gives the **analysis of variance (ANOVA)** decomposition

$$\|y\|^2 = \|\widehat{y}_0\|^2 + \sum_{r=1}^{R} \|\widehat{y}_r - \widehat{y}_{r-1}\|^2 + \|y - \widehat{y}\|^2.$$

Note to Lemma 2

- (i) Let $\mathcal{V}_0 \subset \cdots \subset \mathcal{V}_R$ denote the linear spaces onto which $\mathbb{R}^n$ is projected by $H_0, \ldots, H_R = H$, and suppose that $r \le s$. Now $H_r y \in \mathcal{V}_r$ for any $y \in \mathbb{R}^n$, so as $\mathcal{V}_r \subset \mathcal{V}_s$, $H_r y \in \mathcal{V}_s$. Hence $H_s H_r y = H_r y$ for any $y \in \mathbb{R}^n$, so $H_s H_r = H_r$. This implies that

$$H_s H_r = H_r = H_r^{\mathrm{T}} = (H_s H_r)^{\mathrm{T}} = H_r^{\mathrm{T}} H_s^{\mathrm{T}} = H_r H_s, \quad s \ge r.$$

- (ii) For $r = 1, \ldots, R$, (i) yields $H_0 P_r = H_0 H_r - H_0 H_{r-1} = H_0 - H_0 = 0$, and $H_0 P_{R+1} = H_0(I - H_R) = 0$.
- (iii) The matrices $P_1, \ldots, P_R$ are symmetric because

$$P_r^{\mathrm{T}} = (H_r - H_{r-1})^{\mathrm{T}} = H_r^{\mathrm{T}} - H_{r-1}^{\mathrm{T}} = H_r - H_{r-1} = P_r,$$

  and idempotent because (i) gives

$$\begin{aligned} P_r^2 &= (H_r - H_{r-1})(H_r - H_{r-1}) \\ &= H_r^2 - H_r H_{r-1} - H_{r-1} H_r + H_{r-1}^2 \\ &= H_r - H_{r-1} - H_{r-1} + H_{r-1} \\ &= H_r - H_{r-1} = P_r. \end{aligned}$$

  Moreover if $r < s \le R$, then

$$\begin{aligned} P_r P_s &= (H_r - H_{r-1})(H_s - H_{s-1}) \\ &= H_r H_s - H_r H_{s-1} - H_s H_{r-1} + H_{r-1} H_{s-1} \\ &= H_r - H_r - H_{r-1} + H_{r-1} \\ &= 0. \end{aligned}$$

  The corresponding results for P_{R+1} are equally easy to check.

Analysis of variance II

- Usually $X_0 = 1_n$; then $\widehat{y}_0 = 1_n(1_n^{\mathrm{T}}1_n)^{-1}1_n^{\mathrm{T}}y = \overline{y}1_n$ and

$$\|y\|^2 - \|\widehat{y}_0\|^2 = \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} \overline{y}^2 = \sum_{j=1}^{n} (y_j - \overline{y})^2,$$

  equals n times the empirical variance of $y_1, \ldots, y_n$. Hence

$$\sum_{j=1}^{n} (y_j - \overline{y})^2 = \|y\|^2 - \|\widehat{y}_0\|^2 = \sum_{r=1}^{R} \|\widehat{y}_r - \widehat{y}_{r-1}\|^2 + \|y - \widehat{y}\|^2$$

  decomposes ('analyses') the variability of y around its average $\overline{y}$ into
  - the contributions $\|\widehat{y}_r - \widehat{y}_{r-1}\|^2$ due to adding the columns of X_r to $X_0, \ldots, X_{r-1}$,
  - the **residual sum of squares** $\|y - \widehat{y}\|^2$ left after fitting $X = (X_0, \ldots, X_R)$.
- Large $\|\widehat{y}_r - \widehat{y}_{r-1}\|^2$ implies that X_r explains a lot of the variation of y even after allowing for that explained by $X_0, \ldots, X_{r-1}$.
- The
  - **degrees of freedom** of a fit is the rank ν_r of the corresponding H_r, and the
  - **residual degrees of freedom** is $n - \nu_R = n - p$.

Terms

- A constant column $X_0 = 1_n$ is almost always present in the design matrix, so

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

  where the matrices $X_1, \ldots, X_R$, the **terms**, are successively included.
- The baseline model with only 1_n 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 the baseline we ask which terms lead to large reductions in the residual sum of squares, i.e., best explain the variation of y.
- The successive residual degrees of freedom, i.e., the ranks of the matrices $I - H_r$, are

$$n - 1 = n - \nu_0 \geq n - \nu_1 \geq \cdots \geq n - \nu_R.$$

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

Model formulae

☐ A mean vector such as $1_n\beta_0 + X_1\beta_1 + X_2\beta_2$ is often written as the right-hand side of

$$\texttt{y} \sim \texttt{X1} + \texttt{X2}$$

where

- the columns of 1s is (silently) included first by default,
- `X1` and `X2` represent the vector subspaces of $\mathbb{R}^n$ generated by the corresponding terms, and
- $+$ represents addition of vector subspaces.

☐ Software generally drops any column of a design matrix that is linearly dependent on previous columns, and this affects which elements of β can be estimated and the meaning of estimates corresponding to later columns.

☐ Carefully choosing the order of terms in a model can give easily interpreted estimates of the parameters of interest — for example, if X_2 is full-rank and a column of 1s lies in $\mathrm{span}(X_1) + \mathrm{span}(X_2)$ then

$$\texttt{y} \sim \texttt{X1} + \texttt{X2}, \qquad \texttt{y} \sim \texttt{X2} + \texttt{X1} - \texttt{1},$$

span the same linear space but the second estimates the parameters of β_2 (unadjusted for the mean) and the parameters of β_1, adjusted for the presence of X_2.

ANOVA

Terms	Residual df	Residual SS	Term added	Reduction in residual df	Reduction in SS	Mean square
1_n	$n - \nu_0 = n - 1$	SS_0				
$1_n, X_1$	$n - \nu_1$	SS_1	X_1	$\nu_1 - \nu_0$	$\mathrm{SS}_0 - \mathrm{SS}_1$	$\frac{\mathrm{SS}_0 - \mathrm{SS}_1}{\nu_1 - \nu_0}$
$1_n, X_1, X_2$	$n - \nu_2$	SS_2	X_2	$\nu_2 - \nu_1$	$\mathrm{SS}_1 - \mathrm{SS}_2$	$\frac{\mathrm{SS}_1 - \mathrm{SS}_2}{\nu_2 - \nu_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$1_n, X_1, \ldots, X_R$	$n - \nu_R = n - p$	SS_R	X_R	$\nu_R - \nu_{R-1}$	$\mathrm{SS}_{R-1} - \mathrm{SS}_R$	$\frac{\mathrm{SS}_{R-1} - \mathrm{SS}_R}{\nu_R - \nu_{R-1}}$

☐ The sum of squares when including terms $1_n, X_1, \ldots, X_r$ is

$$\mathrm{SS}_r = \|y - \widehat{y}_r\|^2.$$

☐ The 'mean square' for term X_r,

$$\mathrm{MS}_r = \frac{\mathrm{SS}_{r-1} - \mathrm{SS}_r}{\nu_r - \nu_{r-1}}$$

is the average reduction in SS_r per degree of freedom when X_r is added to the model.

☐ Usually show only the RHS of the table and the bottom line of its LHS (next slide).

ANOVA table

Term added	df	Reduction in SS	Mean square
X_1	$\nu_1 - \nu_0$	$\mathrm{SS}_0 - \mathrm{SS}_1$	$\mathrm{MS}_1 = (\mathrm{SS}_0 - \mathrm{SS}_1)/(\nu_1 - \nu_0)$
X_2	$\nu_2 - \nu_1$	$\mathrm{SS}_1 - \mathrm{SS}_2$	$\mathrm{MS}_2 = (\mathrm{SS}_1 - \mathrm{SS}_2)/(\nu_2 - \nu_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_R	$\nu_R - \nu_{R-1}$	$\mathrm{SS}_{R-1} - \mathrm{SS}_R$	$\mathrm{MS}_R = (\mathrm{SS}_{R-1} - \mathrm{SS}_R)/(\nu_R - \nu_{R-1})$
Residual	$n - \nu_R$	SS_R	$\mathrm{MS}_{\mathrm{Res}} = \mathrm{SS}_R/(n - \nu_R)$

- Used to screen which terms give the largest reductions, comparing MS_r with the residual mean square $\mathrm{MS}_{\mathrm{Res}}$.
- Judge 'significance' of reductions relative to residual using F-tests (later).
- Problem: the order of adding terms matters, so there is no unique reduction in general.

Coefficient of determination

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

$$R^2 = \frac{\|\widehat{y} - \overline{y}1_n\|^2}{\|y - \overline{y}1_n\|^2} = \frac{\{(H - H_0)y)\}^{\mathrm{T}}(H - H_0)y}{\{(I - H_0)y)\}^{\mathrm{T}}(I - H_0)y} = \frac{y^{\mathrm{T}}(H - H_0)y}{y^{\mathrm{T}}(I - H_0)y},$$

  where H_0 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

$$(I_n - H_0)y = (I_n - H)y + (H - H_0)y$$

  are orthogonal (check this).
- Adding columns to X must increase R^2, unlike the **adjusted** R^2,

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

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

$$R^2_0 = \frac{\widehat{y}^{\mathrm{T}}\widehat{y}}{y^{\mathrm{T}}y}, \quad R^2_{0,a} = R^2_0 + (1 - R^2_0)\frac{n}{n-p}.$$

Comments

- ☐ We have supposed that $X_{n\times p}$ has rank p:
  - if X is rank-deficient, then a least squares algorithm usually drops columns that lie in the span of preceding ones, but care is needed to construct X so that the resulting $\widehat{\beta}$ is easy to interpret;
  - if X is nearly rank-deficient, then regularisation may be needed. More later ...
- ☐ Everything so far as purely numerical:
  - least squares estimation is a numerical technique for using X to approximate y;
  - $\widehat{y} = X\widehat{\beta}$ is the resulting approximation, which lies in $\mathrm{span}(X)$;
  - $\widehat{\beta}$ gives the coefficients of the columns of X for the best approximation;
  - the coefficient of determination R^2 measures how much of the overall variation of y was explained by X; and
  - the ANOVA decomposition summarises how much of the variation in y is explained by different subsets of columns of X (terms).
- ☐ For statistics we need to add some distributional assumptions ... shortly ...
- ☐ First some reminders ...

1.2 Inference

Reminder: Moment-generating function

Definition 3 *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.

Reminder: Multivariate normal distribution

A random variable $Y_{n\times 1}$ with real components has the **multivariate normal distribution,** $Y \sim \mathcal{N}_n(\mu, \Omega)$, if $a^{\mathrm{T}}Y \sim \mathcal{N}(a^{\mathrm{T}}\mu, a^{\mathrm{T}}\Omega a)$ for every constant vector $a_{n\times 1}$, and then

(a) Ω is symmetric semi-positive definite with real components and

$$\mathrm{E}(Y) = \mu_{n\times 1}, \quad \mathrm{var}(Y) = \Omega_{n\times n}, \quad M_Y(t) = \exp(t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t), \quad t \in \mathbb{R}^n,$$

where we call μ the **mean vector** and Ω the **(co)variance matrix** of X;

(b) for any constants $a_{m\times 1}$ and $B_{m\times n}$, $a + BY \sim \mathcal{N}_m\left(a + B\mu, B\Omega B^{\mathrm{T}}\right)$;

(c) if $Y^{\mathrm{T}} = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}})$, where Y_1 is $m \times 1$, and μ and Ω are partitioned correspondingly, then the marginal and conditional distributions of Y_1 are also multivariate normal:

$$Y_1 \sim \mathcal{N}_m(\mu_1, \Omega_{11}), \quad Y_1 \mid Y_2 = y_2 \sim \mathcal{N}_m\left\{\mu_1 + \Omega_{12}\Omega_{22}^{-1}(y_2 - \mu_2), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right\};$$

(d) $Y_1 \perp\!\!\!\perp Y_2$ iff $\Omega_{12} = 0$, and $a + BY \perp\!\!\!\perp c + DY$ iff $B\Omega D^{\mathrm{T}} = 0$;

(e) if $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $Y_{n\times 1} \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$; and finally,

(f) Y has a density on $\mathbb{R}^n$ iff Ω is positive definite (i.e., has rank n), and then

$$f(y; \mu, \Omega) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^{\mathrm{T}}\Omega^{-1}(y-\mu)\right\}, \quad y \in \mathbb{R}^n. \tag{1}$$

Note: Multivariate normal distribution

(a) Let e_j denote the n-vector with 1 in the jth 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 $\mathrm{cov}(Y_j, Y_k) = \omega_{jk} = \omega_{kj}$. This gives the mean and covariance matrix of Y.

☐ 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.

(b) 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. Hence linear combinations of normal variables are themselves normal.

(c) Write $Y^{\mathrm{T}} = (Y_1^{\mathrm{T}}, Y_2^{\mathrm{T}})$ and partition μ and Ω conformally. Then

$$M_Y(t) = \exp\left\{t_1^{\mathrm{T}}\mu_1 + t_2^{\mathrm{T}}\mu_2 + \tfrac{1}{2}\left(t_1^{\mathrm{T}}\Omega_{11}t_1 + 2t_1^{\mathrm{T}}\Omega_{12}t_2 + t_2^{\mathrm{T}}\Omega_{22}t_2\right)\right\}$$

and by setting $t_2 = 0$ and $t_1 = 0$ we see that $M_{Y_1}(t_1) = \exp\left(t_1^{\mathrm{T}}\mu_1 + \frac{1}{2}t_1^{\mathrm{T}}\Omega_{11}t_1\right)$ and $M_{Y_2}(t_2) = \exp\left(t_2^{\mathrm{T}}\mu_2 + \frac{1}{2}t_2^{\mathrm{T}}\Omega_{22}t_2\right)$. Hence the marginal distribution of Y_1 is $\mathcal{N}_m(\mu_1, \Omega_{11})$.
For the conditional distribution, note that $W = Y_1 - \Omega_{12}\Omega_{22}^{-1}Y_2$ is a linear combination of Y and

$$\mathrm{E}(W) = \mu_1 - \Omega_{12}\Omega_{22}^{-1}\mu_2, \quad \mathrm{var}(W) = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}, \quad \mathrm{cov}(W, Y_2) = \Omega_{12} - \Omega_{12}\Omega_{22}^{-1}\Omega_{22} = 0.$$

Hence $W \perp\!\!\!\perp Y_2$. As $Y_1 = W + \Omega_{12}\Omega_{22}^{-1}Y_2$ and conditioning on Y_2 does not change the distribution of W,

$$\mathrm{E}(Y_1 \mid Y_2 = y_2) = \mathrm{E}(W) + \Omega_{12}\Omega_{22}^{-1}y_2, \quad \mathrm{var}(Y_1 \mid Y_2 = y_2) = \mathrm{var}(W + \Omega_{12}\Omega_{22}^{-1}y_2) = \mathrm{var}(W).$$

Putting the pieces together gives the stated conditional distribution.

(d) The joint MGF given in (c) factorises iff the variables are independent, and on inspecting it we see that

$$M_Y(t) = M_{Y_1}(t_1)M_{Y_2}(t_2) \quad \Longleftrightarrow \quad \Omega_{12} = 0.$$

The variance matrix of

$$\begin{pmatrix} a \\ c \end{pmatrix} + \begin{pmatrix} B \\ D \end{pmatrix} Y$$

is

$$\begin{pmatrix} B\Omega B^{\mathrm{T}} & B\Omega D^{\mathrm{T}} \\ D\Omega B^{\mathrm{T}} & D\Omega D^{\mathrm{T}} \end{pmatrix},$$

so $a + BY \perp\!\!\!\perp c + DY$ iff $B\Omega D^{\mathrm{T}} = 0$.

Note: Multivariate normal distribution II

(e) 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 $\sum_{j=1}^n u_j\mu = u^{\mathrm{T}}\mu 1_n$ and variance $\sum_{j=1}^n u_j^2\sigma^2 = u^{\mathrm{T}}\sigma^2 I_n u$, so $Y \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, as required.

(f) Since Ω is symmetric and positive semi-definite, the spectral theorem tells us that we may write $\Omega = ADA^{\mathrm{T}}$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$ contains the (real) eigenvalues of Ω, with $d_1 \geq \cdots \geq d_n \geq 0$, and A is a $n \times n$ orthogonal matrix, i.e., $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I_n$ and $|A| = 1$. The columns $A_1, \ldots, A_n$ of A are the eigenvectors corresponding to the respective eigenvalues,

$$\Omega = ADA^{\mathrm{T}} = \sum_{j=1}^n d_j a_j a_j^{\mathrm{T}},$$

with $|\Omega| = |ADA^{\mathrm{T}}| = |A| \times |D| \times |A^{\mathrm{T}}| = |D|$ and $\Omega^{-1} = AD^{-1}A^{\mathrm{T}}$ if the inverse exists.

☐ Now let $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$, and $u \in \mathbb{R}^n$, set and consider

$$u^{\mathrm{T}}(\mu + AD^{1/2}Z) = u^{\mathrm{T}}\mu + \sum_{j=1}^n Z_j u^{\mathrm{T}} a_j d_j^{1/2}.$$

This is a linear combination of normal variables, so it has a normal distribution, with mean $u^{\mathrm{T}}\mu$ and variance

$$\mathrm{var}\left(u^{\mathrm{T}}\mu + \sum_{j=1}^n Z_j u^{\mathrm{T}} a_j d_j^{1/2}\right) = \sum_{j=1}^n d_j (u^{\mathrm{T}}a_j)^2 \mathrm{var}(Z_j) = u^{\mathrm{T}}\left(\sum_{j=1}^n d_j a_j a_j^{\mathrm{T}}\right) u = u^{\mathrm{T}}\Omega u,$$

so we can write $X = \mu + AD^{1/2}Z \sim N_n(\mu, \Omega)$.

☐ If Ω has rank n, then $d_n > 0$. The change of variables $z \mapsto x = \mu + AD^{1/2}z$ has Jacobian

$$\left|\frac{\partial x}{\partial z}\right| = |AD^{1/2}| = |A||D|^{1/2} = 1 \times |D|^{1/2} = |\Omega|^{1/2} > 0.$$

Moreover $z = D^{-1/2}A^{\mathrm{T}}(x - \mu)$, and therefore $z^{\mathrm{T}}z = (x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)$. Hence using the joint density of Z, $f_Z(z) = (2\pi)^{-n/2}\exp(-\sum_{j=1}^n z_j^2/2)$,

$$f_X(x) = f_Z(z)|_{z=D^{-1/2}A^{\mathrm{T}}(x-\mu)} \left|\frac{\partial z}{\partial x}\right| = (2\pi)^{-n/2} \exp\left(-\frac{z^{\mathrm{T}}z}{2}\right)\Bigg|_{z=D^{-1/2}A^{\mathrm{T}}(x-\mu)} |\Omega|^{-1/2},$$

which reduces to (1). If $d_n = 0$, then the Jacobian is zero, so the transformation $z \mapsto x$ is singular and X does not have a density on $\mathbb{R}^n$.

☐ Now suppose that $d_m > d_{m+1} = 0$, so just m eigenvalues of Ω are positive. Then

$$X = \mu + \sum_{j=1}^m Z_j a_j d_j^{1/2} \in \mathcal{S} = \mu + \mathrm{span}(a_1, \ldots, a_m),$$

where $\mathcal{S}$ is a hyperplane of dimension m passing through μ and generated by the vectors $a_1, \ldots, a_m$. In this case the previous argument shows that X has an m-dimensional Gaussian density on $\mathcal{S}$, but places no probability elsewhere.

Reminder: χ^2 distribution

Definition 4 *If* $Y_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, *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** $\delta^2 = (\mu_1^2 + \cdots + \mu_\nu^2)/\sigma^2$; *we write* $W \sim \sigma^2 \chi^2_\nu(\delta^2)$. *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* $W \sim \chi^2_\nu$, *its MGF is* $M_W(t) = (1-2t)^{-\nu/2}$, *and its* p*-quantile is* $c_\nu(p)$.

Chi-square variables satisfy

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

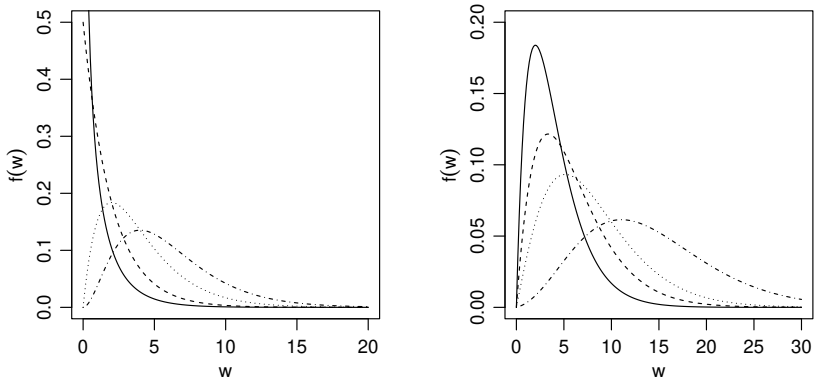
$$f(w) = \frac{\beta^\alpha w^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta w}, \quad w > 0, \quad \alpha, \beta > 0,$$

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

Reminder: χ^2_ν densities

Left: central densities with $\nu = 1, 2, 4, 6$ (solid, large dashes, small dashes, dot-dash).

Right: non-central densities with $\nu = 4$ and $\delta = 0, 2, 4, 10$ (solid, large dashes, small dashes, dot-dash).



Reminder: Student t distribution

Definition 5 *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 $\nu \geq 2$ and $\nu \geq 3$ respectively, and then

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

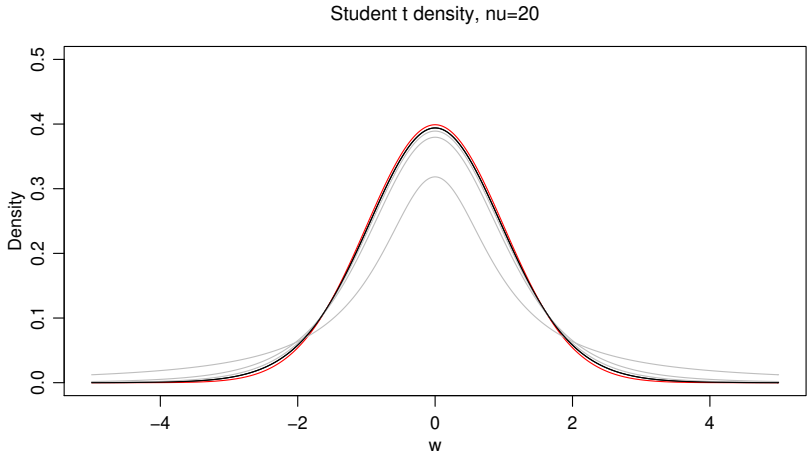
- with $\nu = 1$ we have the **Cauchy density**,

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

  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).

Reminder: Student t densities

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



Reminder: F distribution

Definition 6 *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* $F_{\nu_1,\nu_2}(p)$.

Reminder: 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

- ☐ Least squares fitting gives a deterministic description of the variation in some numbers y in terms of other numbers X.
- ☐ A **statistical model** is a description of data y in terms of a collection of probability distributions on the sample space for y.
- ☐ 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 are in general use for the linear model:
  - **second-order assumptions**,

$$y \sim (X\beta, \sigma^2 V), \quad \text{i.e.,} \quad \mathrm{E}(y) = X\beta, \quad \mathrm{var}(y) = \sigma^2 V_{n\times n};$$

  - **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 positive definite (co)variance matrix $\sigma^2 V$.
- 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 7 *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 8 (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.*

- Obviously these results hold under the (stronger) normal assumptions.
- The Gauss–Markov theorem only concerns linear estimators. Nonlinear estimators of β might have smaller variance than $\sigma^2 (X^{\mathrm{T}} X)^{-1}$ (and in fact the optimal maximum likelihood estimators of β for non-normal models will be nonlinear in y).

Note to Lemma 7

- ☐ Recall that expectation is linear, and that $\mathrm{var}(A_{p\times n}y) = A\mathrm{var}(y)A^{\mathrm{T}}$.
- ☐ 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 $\mathrm{E}(yy^{\mathrm{T}}) = \mathrm{var}(y) + \mathrm{E}(y)\mathrm{E}(y)^{\mathrm{T}} = \sigma^2 I_n + X\beta\beta^{\mathrm{T}}X^{\mathrm{T}}$, 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 $\mathrm{E}(S^2) = \sigma^2$, because

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

Note to Theorem 8

- ☐ 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.

- ☐ This is a finite-sample result that holds for all n and X (of rank p, with $n \geq p$).

Second-order assumptions and large samples

☐ We can write $y_j = x_j^{\mathrm{T}}\beta + \sigma\varepsilon_j$, where $\varepsilon_j \overset{\text{ind}}{\sim} (0,1)$, so

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

say, where $a_1, \ldots, a_n$ are $p \times 1$ vectors. We have $\mathrm{E}(\widehat{\beta}) = \beta$ and $\mathrm{var}(\widehat{\beta}) = (X^{\mathrm{T}}X)^{-1}$, but is $\widehat{\beta}$ approximately normal for large n?

☐ The a_j, or equivalently X, must be such that no single y_j can dominate in $n^{-1}\sum a_j\varepsilon_j$.

Theorem 9 (no proof) *Let $\{X_n\}$ be a sequence of $n \times p$ design matrices each of rank p, let $h_{11}^n, \ldots, h_{nn}^n$ be the diagonal elements of the hat matrices $X_n(X_n^{\mathrm{T}}X_n)^{-1}X_n^{\mathrm{T}}$ and let $y_n \sim (X_n\beta, \sigma^2 I_n)$ for each n. If*

$$\lim_{n\to\infty}\max_{j=1,\ldots,n} h_{jj} = 0,$$

then the corresponding sequence of least squares estimators $\widehat{\beta}_n$ satisfies

$$(X_n^{\mathrm{T}}X_n)^{1/2}(\widehat{\beta}_n - \beta) \xrightarrow{D} \mathcal{N}_p(0, \sigma^2 I_p), \quad n \to \infty,$$

i.e., if H has a 'well-behaved' diagonal, then $\widehat{\beta} \dot{\sim} \mathcal{N}_p\{\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}\}$ in large samples.

Normal-theory linear model

The following results allow exact inferences for β and σ^2, and in analysis of variance.

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

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

Lemma 11 *If $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$ and H is symmetric and idempotent with rank p, then $y^{\mathrm{T}}Hy \sim \sigma^2\chi_p^2(\delta^2)$, where $\sigma^2\delta^2 = \mu^{\mathrm{T}}H\mu$.*

Theorem 12 *If $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$ and a linear model is fitted whose design matrix X is structured as in Lemma 2, then the sums of squares in the ANOVA decomposition*

$$\sum_{j=1}^{n}(y_j - \overline{y})^2 = \sum_{r=1}^{R}\|\widehat{y}_r - \widehat{y}_{r-1}\|^2 + \|y - \widehat{y}\|^2 = \sum_{r=1}^{R+1}\|P_r y\|^2$$

are independent and $\|P_r y\|^2 \sim \sigma^2\chi^2_{\nu_{r-1}-\nu_r}(\delta_r^2/\sigma^2)$, where $\sigma^2\delta_r^2 = \mu^{\mathrm{T}}P_r\mu$. If X_r does not explain any variation in μ after allowing for $X_0, \ldots, X_{r-1}$, then $P_r\mu = 0$, so $\delta_r^2 = 0$.

Theorem 12 implies that the sums of squares for terms that explain variation in y will tend to be larger than sums of squares for other terms, which can be used to estimate σ^2.

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 7.
- ☐ 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 $D = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$ and $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_n$. 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 Lemma 11

The spectral decomposition of H is UDU^{T}, where D is diagonal with p 1's and $n - p$ 0's, and $Z = U^{\mathrm{T}}y \sim \mathcal{N}_n(U^{\mathrm{T}}\mu, \sigma^2 I_n)$; note that the Z_j are independent. Now

$$y^{\mathrm{T}}Hy = (U^{\mathrm{T}}y)^{\mathrm{T}}D(U^{\mathrm{T}}y) = \sum_{j=1}^{n} d_j Z_j^2 = \sum_{j:d_j=1} Z_j^2,$$

which has a (possibly non-central) χ^2 distribution with $p = \mathrm{tr}(H)$ degrees of freedom, scale parameter σ^2 and

$$\sigma^2\delta^2 = \sum_{j:d_j=1} \mathrm{E}(Z_j)^2 = \sum_{j=1}^{n} d_j \mathrm{E}(Z_j)^2 = (U^{\mathrm{T}}\mu)^{\mathrm{T}}D(U^{\mathrm{T}}\mu) = \mu^{\mathrm{T}}H\mu.$$

Note to Theorem 12

- As $P_r P_s = 0$ for $r \neq s$, we have $\mathrm{cov}(P_r y, P_s y) = P_r \mathrm{var}(y) P_s^{\mathrm{T}} = \sigma^2 P_r P_s = 0$, i.e., $P_r y$ and $P_s y$ are independent. Hence the terms in the ANOVA decomposition are independent.
- P_r is a symmetric idempotent matrix, so Lemma 11 gives

$$\|P_r y\|^2 \sim \sigma^2 \chi^2_\nu(\delta_r^2/\sigma^2), \quad \delta_r^2 = \mu^{\mathrm{T}} P_r \mu,$$

  where $\nu = \mathrm{rank}(P_r)$. These ranks are $\nu_{r-1} - \nu_r$ for $r = 1, \ldots, R$, and $\nu_{R+1} = n - p$ for $P_{R+1} = I_n - H$.
- If X_r does not explain any variation in μ after allowing for $X_0, \ldots, X_{r-1}$, then $H_r \mu = H_{r-1}\mu \in \mathcal{V}_{r-1}$, i.e., $P_r \mu = 0$, and thus $\delta_r^2 = 0$.

Inference on β

- Theorem 10 implies that for any constant $c_{p\times 1}$, $c^{\mathrm{T}}\widehat{\beta} \sim \mathcal{N}\{c^{\mathrm{T}}\beta, \sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c\}$, so

$$Z = \frac{c^{\mathrm{T}}\widehat{\beta} - c^{\mathrm{T}}\beta}{\sigma\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \sim \mathcal{N}(0,1) \quad \perp\!\!\!\perp \quad (n-p)S^2/\sigma^2 = W \sim \chi^2_{n-p},$$

  and thus

$$\frac{c^{\mathrm{T}}\widehat{\beta}_r - c^{\mathrm{T}}\beta_r}{S\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} = \frac{Z}{\sqrt{W/(n-p)}} \sim t_{n-p}.$$

- Let v_{rs} denote the (r,s) element of $(X^{\mathrm{T}}X)^{-1}$, so v_{rr} denotes its rth diagonal element.
- Different choices of c allow inferences on the elements of β.
- For example, if $c^{\mathrm{T}} = (c_1, \ldots, c_p)$, $c_r = 1$ and $c_s = 0$ for $s \neq r$, then $c^{\mathrm{T}}\beta = \beta_r$, and we
  - 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

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

- Likewise we can compare β_r and β_s by setting $c_r = 1$, $c_s = -1$ and all other $c_t = 0$.

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?)

1.3 Analysis of Variance slide 33

Analysis of variance

- We previously saw that

$$\sum_{j=1}^{n}(y_j-\overline{y})^2 = \|y\|^2 - \|\widehat{y}_0\|^2 = \sum_{r=1}^{R}\|\widehat{y}_r-\widehat{y}_{r-1}\|^2 + \|y-\widehat{y}\|^2$$

  decomposes ('analyses') the variability of y around its average $\overline{y}$ into
  - the contributions $\|\widehat{y}_r-\widehat{y}_{r-1}\|^2$ due to adding the columns of X_r to $X_0,\ldots,X_{r-1}$,
  - the **residual sum of squares** $\|y-\widehat{y}\|^2$ left after fitting $X=(X_0,\ldots,X_R)$.
- Large $\|\widehat{y}_r-\widehat{y}_{r-1}\|^2$ implies that X_r explains a lot of the variation of y even after allowing for that explained by $X_0,\ldots,X_{r-1}$.
- Theorem 12 implies that under the normal assumptions, and if $\mathrm{E}(y)=\mu$ lies in the column space of X, the sums of squares on the RHS above are independent and satisfy

$$\|\widehat{y}_r-\widehat{y}_{r-1}\|^2 = \|P_r y\|^2 \sim \sigma^2\chi^2_{\nu_{r-1}-\nu_r}(\delta_r^2/\sigma^2) \quad \perp\!\!\!\perp \quad \|y-\widehat{y}\|^2 \sim \sigma^2\chi^2_{n-p}.$$

  Hence if $\delta_r^2=0$, i.e., $\mu\in\mathrm{span}(X_0,\ldots,X_{r-1})$, then

$$\frac{\|\widehat{y}_r-\widehat{y}_{r-1}\|^2/(\nu_{r-1}-\nu_r)}{\|y-\widehat{y}\|^2/(n-p)} \sim F_{\nu_{r-1}-\nu_r,n-p}.$$

ANOVA table

Term added	df	Reduction in SS	Mean square
X_1	$n-1-\nu_1$	$\mathrm{SS}_0-\mathrm{SS}_1$	$\mathrm{MS}_1=(\mathrm{SS}_0-\mathrm{SS}_1)/(n-1-\nu_1)$
X_2	$\nu_1-\nu_2$	$\mathrm{SS}_1-\mathrm{SS}_2$	$\mathrm{MS}_2=(\mathrm{SS}_1-\mathrm{SS}_2)/(\nu_1-\nu_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_R	$\nu_{R-1}-\nu_R$	$\mathrm{SS}_{R-1}-\mathrm{SS}_R$	$\mathrm{MS}_R=(\mathrm{SS}_{R-1}-\mathrm{SS}_R)/(\nu_{R-1}-\nu_R)$
Residual	$\nu_R=n-p$	SS_R	$\mathrm{MS}_{\mathrm{Res}}=\mathrm{SS}_R/\nu_R$

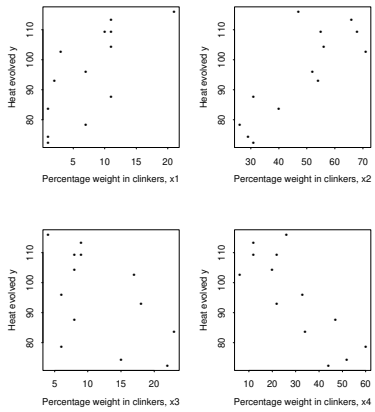
- If $\mu\in\mathrm{span}(X)$ then the residual mean square $\mathrm{MS}_{\mathrm{Res}}$ gives an estimate of σ^2.
- We test for an effect of term X_r by noting that
  - if X_r explains no more than $(X_0,\ldots,X_{r-1})$, then

$$F_r=\frac{\mathrm{MS}_r}{\mathrm{MS}_{\mathrm{Res}}} \sim F_{\nu_{r-1}-\nu_r,\nu_R},$$

  - if X_r does have additional explanatory power, then the distribution of MS_r is shifted to the right, and we expect F_r to be large relative to its null distribution.

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, the ANOVA and ANOVA table depend on the order of inclusion of terms.
- ☐ Its interpretation is unclear if X_r is significant when included early, but not when it is included late. Is the term 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 the reductions do not depend on the order of inclusion. Hooray!

Balance

- ☐ Balanced design matrices induce orthogonality after fitting 1_n (or a more complex design X_0).
- ☐ Gram–Schmidt orthogonalisation with respect to early terms makes later terms mutually orthogonal, leading to a clear interpretation of the ANOVA for the later terms.
- ☐ If we write $H_0 = X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}$ and let

$$Z_r = P_0X_r = (I_n - H_0)X_r, \quad r = 1, 2,$$

denote the versions of X_1 and X_2 after adjusting for X_0, then

$$\begin{aligned} X_0\beta_0 + X_1\beta_1 + X_2\beta_2 &= (X_0\beta_0 + H_0X_1\beta_1 + H_0X_2\beta_2) + P_0X_1\beta_1 + P_0X_2\beta_2 \\ &= Z_0\gamma_0 + Z_1\beta_1 + Z_2\beta_2, \end{aligned}$$

say, and $Z_1^{\mathrm{T}}Z_0 = Z_2^{\mathrm{T}}Z_0 = 0$, because $P_0X_0 = P_0H_0 = 0$.

- ☐ If the design satisfies $Z_1^{\mathrm{T}}Z_2 = 0$, then the order of inclusion of X_1, X_2 is irrelevant, provided X_0 is already present in the fit.

Example 13 (3×2 layout) *Observations and their means written as*

$$\begin{array}{cc} y_{11} & y_{12} \\ y_{21} & y_{22}, \\ y_{31} & y_{32} \end{array} \qquad \begin{array}{cc} \mu + \alpha_1 + \delta_1 & \mu + \alpha_2 + \delta_1 \\ \mu + \alpha_1 + \delta_2 & \mu + \alpha_2 + \delta_2. \\ \mu + \alpha_1 + \delta_3 & \mu + \alpha_2 + \delta_3 \end{array}$$

Note to Example 13

☐ In terms of the parameter vector $(\mu, \alpha_1, \alpha_2, \delta_1, \delta_2, \delta_3)^{\mathrm{T}}$, the design matrix is

$$X^*_{6\times 6} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \quad \text{with responses} \quad y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix},$$

with $X_0 \equiv 1_6$ the first column of X^*, columns 2–3 the term X^*_1 for columns, and columns 4–6 the term X^*_2 for rows.

☐ This model has six parameters, but they cannot all be estimated, because X_0 lies in the column spaces of X^*_1 and X^*_2, and it is easy to check that X^* has rank 4. The usual way to deal with this is to set $\alpha_1 = \delta_1 = 0$, corresponding to dropping columns 2 and 4 of X^*, giving the so-called *corner-point parametrization* in which the means are

$$\begin{array}{cc} y_{11} & y_{12} \\ y_{21} & y_{22}, \\ y_{31} & y_{32} \end{array} \qquad \begin{array}{cc} \mu & \mu + \alpha_2 \\ \mu + \delta_2 & \mu + \alpha_2 + \delta_2, \\ \mu + \delta_3 & \mu + \alpha_2 + \delta_3 \end{array}$$

i.e.,

- the 'grand mean' μ corresponds to the mean of observations with the first level of every factor,
- α_2 corresponds to the mean difference between column 2 and column 1,
- δ_2 corresponds to the mean difference between row 2 and row 1, and
- δ_3 corresponds to the mean difference between row 3 and row 1.

This is the default in R. More rarely we might set $\sum_c \alpha_c = \sum_r \delta_r = 0$.

☐ Even after these columns are dropped to give

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

the terms X_1 for columns and X_2 for rows 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_2 and without α_2. This balance implies that if 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.

1.4 Diagnostics

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.
- Concerns about primary aspects suggest that we should start again.
- Concerns 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

- The **raw residuals**

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

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

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

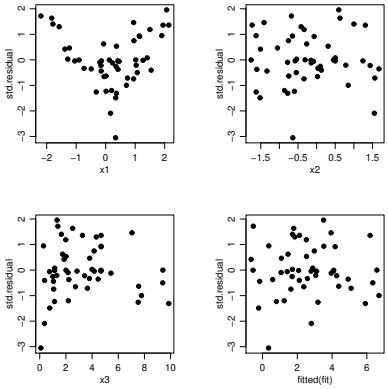
- To (roughly) equalise the variances we define **standardized residuals**

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

  with s replacing σ. Then $\mathrm{E}(r_j) = 0$ and $\mathrm{var}(r_j) \doteq 1$.
- Although $e^{\mathrm{T}}\widehat{y} = \mathrm{cov}(e, \widehat{y}) = 0$ (check!), this only implies no linear relation between e and $\widehat{y}$.
- We check
  - linearity by plotting r_j against the covariates (those in X and those not in X);
  - constant variance by plotting r_j (or $|r_j|$) against fitted values $\widehat{y}_j$;
  - independence by ACF of residuals (if data time-ordered);
  - for outliers, which are visible as unusual residuals; and
  - normality using a 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 $\widehat{y}^{\mathrm{T}} e = 0$):

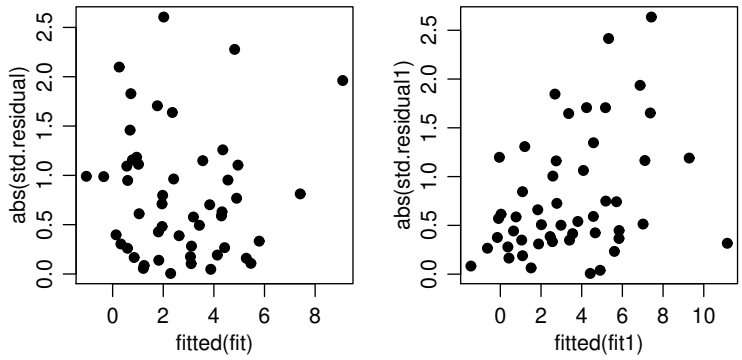


Checking the variance

☐ 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 $\widehat{y}$:



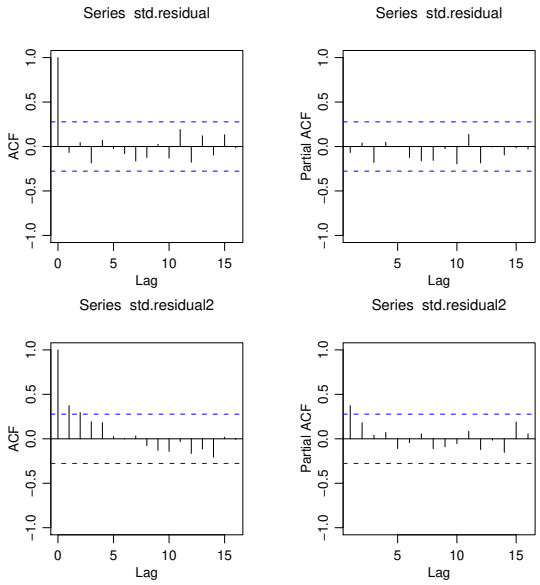
Checking independence

- ☐ 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 for outliers and normality

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

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

against (approximate) expected normal order statistics

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

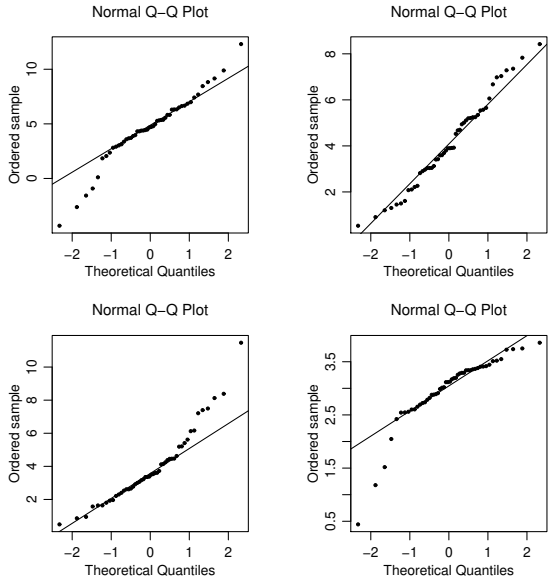
☐ Normality — roughly straight line, slope σ, intercept μ.

☐ Outliers, skewness, heavy tails (easily) seen.

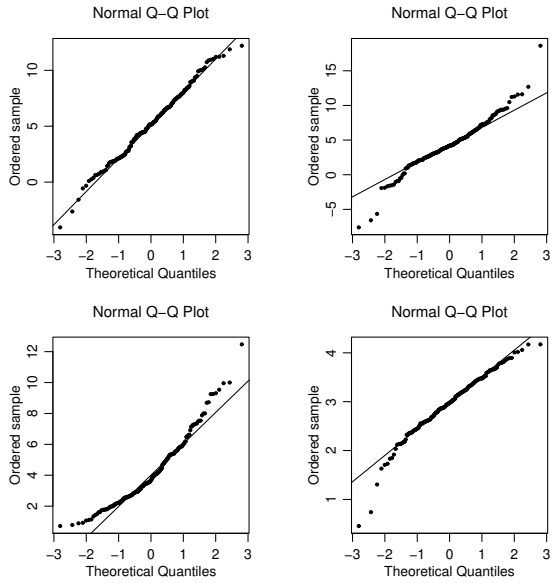
☐ Beware over-interpretation of such plots when n is small — often useful to add simulation envelope.

☐ 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

$$\mathrm{var}(y_j - \widehat{y}_j) = \mathrm{var}(y_j - x_j^{\mathrm{T}}\beta) = \sigma^2(1 - h_{jj}),$$

  having **leverage** $h_{jj} \doteq 1$ implies that $\widehat{y}_j \approx y_j$ — need one parameter to fit this case.
- ☐ As $\mathrm{tr}(H) = \sum_{j=1}^{n} h_{jj} = p$, 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

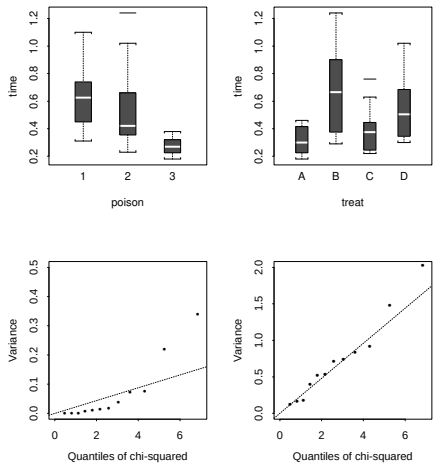
- ☐ Linear model for y may be better applied for some transformation $g(y)$, especially if some y are much larger than others, or the variance is non-constant.
- ☐ Survival times y_{ptj} in 10-hour units of animals in a 3×4 factorial experiment with four replicates, with (below) average (standard deviation) 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:

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

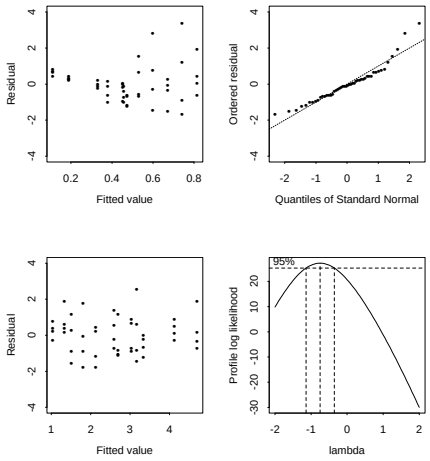
- Analyses of variance with responses y and y^{-1}. For MS and F read 'Mean square' and 'F statistic'.

- The terms explain appreciably more of the variation of y^{-1}, suggesting that this is a preferable choice of response.

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 $1/y$.



Summary on model-checking

- ☐ Recall the distinction between primary and secondary assumptions. Use of the standard linear model when the secondary assumptions fail leads to inefficient estimation and over-confident uncertainty assessment, but is not usually disastrous per se.
- ☐ When they fail ...
  - **Linearity** (primary): add terms (e.g., x^2) to the model, transform the covariate (e.g., to $\log x$), or question the basic setup;
  - **Constant variance** (secondary): use a response transformation (below), weighted least squares, or question primary aspects. Non-constant variance affects uncertainty assessment, but not estimation;
  - **Lack of correlation (independence)** (secondary): use a correlated error model (e.g., time series or random effects). Dependence affects uncertainty assessment, but not estimation;
  - **normality** (secondary): often does not matter, because the CLT applies to the estimators. It does matter for prediction, which is affected by the distribution of individual responses;
- ☐ Checking leverage and influence may be useful in small and moderate samples, but rarely in large samples. In any case, automatic dropping of outlying and/or influential cases is dangerous!

1.5 Model Building

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

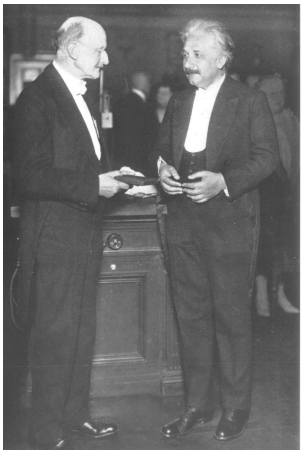
- ☐ **Collect** data intended to answer question of interest;
- ☐ **examine** data (graphs, look for outliers, problems with sampling scheme);
- ☐ **choose/construct** response variable (transformations? independence?);
- ☐ **consider** what models are coherent with context of problem (limiting properties, 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, $\mathrm{AIC_c}$, 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 $y \sim (\mu, \sigma^2 I)$ and fit $X\beta$, obtaining fitted value

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

where $H\mu = \mu$, i.e., $(I - H)\mu = 0$ if $\mu \in \mathrm{span}(X)$, but otherwise $(I - H)\mu \neq 0$.
We have a new data set $y_+ \sim (\mu, \sigma^2 I)$, 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 $e_+ = y_+ - X\widehat{\beta}$ and note that as the trace of a scalar is the scalar and trace is a linear operator,

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

Now as y_+ and y are independent and $\mathrm{var}(X\widehat{\beta}) = \sigma^2 H$, we have

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

so the computation above gives

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

because $\mathrm{tr}(I + H) = n + p$ 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 it depends on 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 $\Delta_{\mathcal{S}}$, so it would be preferable to use the entire dataset ...

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, but the lemma below implies that with one fit we have

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

Lemma 14 *For a fit $\widehat{y} = Hy$ where H has jth diagonal element h_{jj} and $\widehat{y}_{j,-j}$ is the fitted value for y_j obtained when (x_j, y_j) is dropped,*

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

and therefore

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

Note to Lemma 14

- Consider any linear fit $\widehat{y} = Hy$, and note that $\widehat{y}_j = \sum_{i=1}^{n} h_{ji}y_i$.
- Now suppose we leave out (x_j, y_j) and compute the corresponding (penalized) estimate

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

  and fitted value $y_j^* = \widehat{y}_{j,-j} = x_j^{\mathrm{T}}\widehat{\beta}_{-j}$ corresponding to x_j.
- Inserting (x_j, y_j^*) back 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. For this new dataset,

$$y_j^* = \sum_{i\neq j} h_{ji}y_i + h_{jj}y_j^* = \sum_{i=1}^{n} h_{ji}y_i + h_{jj}(y_j^* - y_j) = \widehat{y}_j + h_{jj}(y_j^* - y_j)$$

  so

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

  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.

Generalized cross-validation

- Leave-one-out CV can be unstable if some of the h_{jj} are large.
- **Generalised cross-validation (GCV)** replaces all the h_{jj} by their 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 that these only require the second-order assumptions.

Note: Properties of GCV

We have $(1-p/n)^2\mathrm{GCV} = e^{\mathrm{T}}e$ where $e = y - X\widehat{\beta} = (I-H)y \sim ((I-H)\mu, (I-H)\sigma^2)$, and

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

Now note that $\mathrm{tr}(I-H) = n-p$ 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**

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

  between **candidate model** $f_\theta \equiv f(y;\theta)$ and true model g, based on $Y_1, \ldots, Y_n \overset{\mathrm{iid}}{\sim} g$.
- Suppose that θ_g minimises $D(f_\theta, g)$ within the family of candidate models, and is estimated by the MLE $\widehat{\theta}$, with log likelihood $\widehat{\ell}$.
- We suppose there is an independent sample $Y_1^+, \ldots, Y_n^+ \overset{\mathrm{iid}}{\sim} g$ and aim to estimate

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

  the outer expectation is over the distribution of $\widehat{\theta}$, which is independent of Y^+.
- After tedious expansions we end up trying to minimise the **Akaike information criterion**

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

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

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

  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

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

  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

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

  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:

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

- **Bayes' information criterion**

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

- **Mallows C_p**:

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

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 the model with a constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms improves the fit, stop; otherwise
  3. update the current model to include the most useful new term; go to 1
- **Backward elimination**: starting from the model with all terms,
  1. if all terms are 'useful', stop; otherwise
  2. update current model by dropping the 'least useful' 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
- 'Useful' means 'reduces the AIC' (but in the past meant 'is significant using an F test').

Stepwise methods: Comments

- Original formulation of stepwise used F tests (or even arbitrary numbers!) to assess significance, but this finds spurious models.
- Systematic search minimising AIC or similar over all possible models is preferable, but is often infeasible.
- Compare AICs for different models at each step—i.e., use 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 — especially if the interpretations of competing models differ.

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)	

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:

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

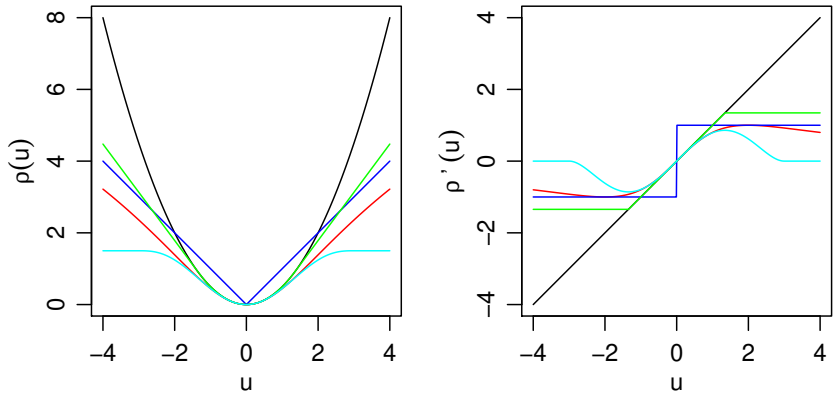
- The function $\rho'(u)$ is also called the **influence function** of the estimator, as its value determines what influence an observation at u has on the estimator:
  - Huber $\rho'(u)$ is bounded,
  - t_ν function is bounded and **redescending**, as $\lim_{u\to\pm\infty} \rho'(u) = 0$;
  - Tukey's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

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

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

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

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 (**??**) 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

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

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 $\rho'(u) = \text{sign}(u)$ 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

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

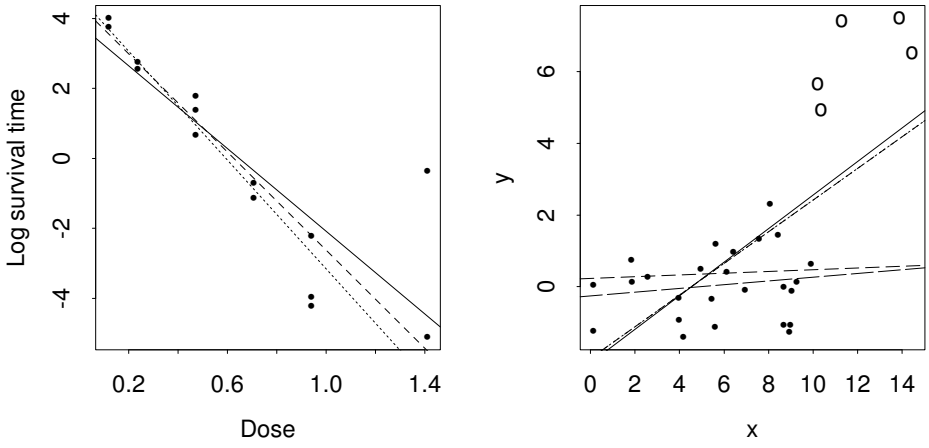
  for example, the Huber estimator is 95% efficient if $c = 1.345$.
- In practice need to balance robustness and efficiency, increasing the latter by increasing c.
- High numbers of outliers can wreck M-estimators.
- Highly robust **least trimmed squares** estimators obtained by minimising

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

  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 85)

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
	No outliers	With outliers	Huber	Biweight	squares
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

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

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 round the cusp of 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).

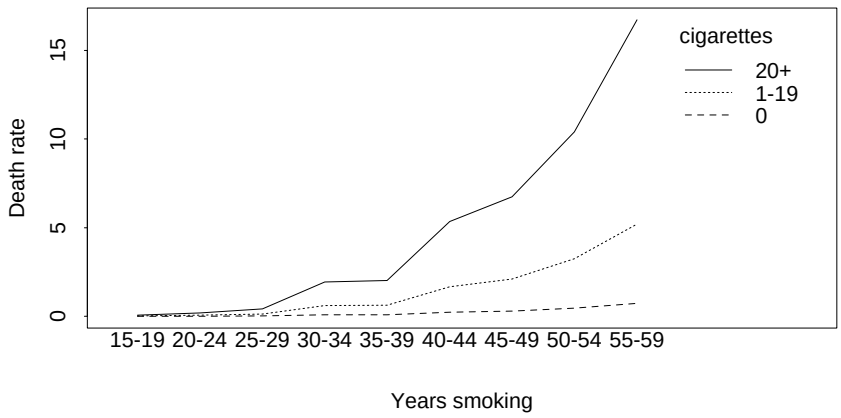
Smoking data

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.

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 $\text{E}(y)$ 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.

Revision: Likelihood

Definition 15 *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*

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad I(\theta) = \mathrm{E}\left\{J(\theta)\right\};$$

these are $p \times p$ matrices if θ has dimension p.

Revision: 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 $\widehat{\theta}$.

Revision: 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

$$|\partial^3 \log f(Y_j;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| \leq g(Y_j)$$

   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.

Revision: 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.

Revision: 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 $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is nested within the model $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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.

Revision: 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 jth observation.
- ☐ Suppose ϕ known (for now), suppress it, and estimate β.

Example 16 (Normal regression model) *Express the normal regression model in the terms above.*

Note to Example 16

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

$$\frac{\partial \ell(\widehat{\beta})}{\partial \beta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} \frac{\partial \ell}{\partial \eta} = \frac{\partial \eta^{\mathrm{T}}}{\partial \beta} u(\widehat{\beta}) = \sum_{j=1}^{n} \frac{\partial \eta_j}{\partial \beta} \frac{\partial \ell_j\{\eta_j(\beta)\}}{\partial \eta_j} = 0, \tag{6}$$

where $u(\beta)$ is $n \times 1$ vector with jth element $\partial \ell / \partial \eta_j$.

IWLS II

- Newton–Raphson update step:

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z,$$

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 (3), to obtain

$$\frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta) + \left\{ \sum_{j=1}^{n} \frac{\partial \eta_j(\beta)}{\partial \beta} \frac{\partial^2 \ell_j(\beta)}{\partial \eta_j^2} \frac{\partial \eta_j(\beta)}{\partial \beta^{\mathrm{T}}} + \sum_{j=1}^{n} \frac{\partial^2 \eta_j(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} u_j(\beta) \right\} (\widehat{\beta} - \beta) \doteq 0. \tag{7}$$

  If we denote the $p \times p$ matrix in braces on the left by $-J(\beta)$, assumed invertible, we can rearrange (**??**) to obtain

$$\widehat{\beta} \doteq \beta + J(\beta)^{-1} \frac{\partial \eta^{\mathrm{T}}(\beta)}{\partial \beta} u(\beta). \tag{8}$$

  This suggests that maximum likelihood estimates may be obtained by starting from a particular β, using (**??**) to obtain $\widehat{\beta}$, then setting β equal to $\widehat{\beta}$, and iterating (**??**) 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

$$I(\beta) = X(\beta)^{\mathrm{T}} W(\beta) X(\beta), \tag{9}$$

  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 (**??**), we obtain

$$\widehat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W (X\beta + W^{-1} u) = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W z, \tag{10}$$

  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 (**??**), or from the profile log likelihood $\ell_{\mathrm{p}}(\phi)$.

Examples

Example 17 (Normal nonlinear model) *Give the components of the IWLS algorithm for the normal nonlinear model.*

Note to Example 17

- 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 (**??**) is

$$\widehat{\beta} \doteq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + y - \eta), \tag{11}$$

where X and η are evaluated at the current β. Here $\eta \neq X\beta$ and (**??**) 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 18 (Normal linear model) *Find the difference of deviances in the normal linear model.*

Note to Example 18

- 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 $q < p$,

$$D_B - D_A = \phi^{-1}\sum_{j=1}^{n}\left\{(y_j - \widehat{\eta}_j^B)^2 - (y_j - \widehat{\eta}_j^A)^2\right\} \quad \dot{\sim} \quad \chi^2_{p-q}$$

  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 denominator degrees of freedom equal to the degrees of freedom for estimation of ϕ.

2.2 Model Checking

slide 108

Model checking

- 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 19 (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 19

- As the data are annual maxima, it is more appropriate to suppose that y_j has the Gumbel density

$$f(y_j;\eta_j,\tau)=\tau^{-1}\exp\left\{-\frac{y_j-\eta_j}{\tau}-\exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\}, \tag{12}$$

  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

$$\ell_j(\eta_j,\tau)=-\log\tau-\frac{y_j-\eta_j}{\tau}-\exp\left(-\frac{y_j-\eta_j}{\tau}\right), \tag{13}$$

  and it is straightforward to establish that

$$\frac{\partial\ell_j(\eta_j,\tau)}{\partial\eta_j}=\tau^{-1}\left\{1-\exp\left(-\frac{y_j-\eta_j}{\tau}\right)\right\},\quad \mathrm{E}\left\{-\frac{\partial^2\ell_j(\eta_j,\tau)}{\partial\eta_j^2}\right\}=\tau^{-2},$$

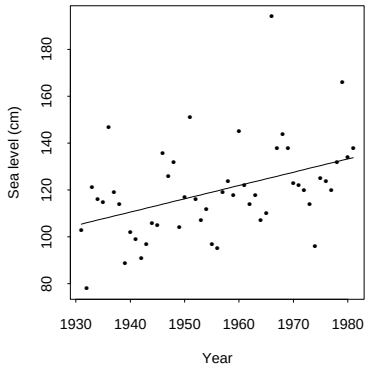
  that $\partial\eta/\partial\beta^{\mathrm{T}}=X$ is the $n\times 2$ matrix whose jth row is $(1,x_j-\overline{x})$, and $W=\tau^{-2}I_n$. Hence (**??**) becomes $\widehat{\beta}\doteq(X^{\mathrm{T}}X)^{-1}(X\beta+\tau^2u)$, where the jth element of u is $\tau^{-1}[1-\exp\{-(y_j-\eta_j)/\tau\}]$.

- Here it is simplest to fix τ, to obtain $\widehat{\beta}$ by iterating (**??**) 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 (**??**) 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 1 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.

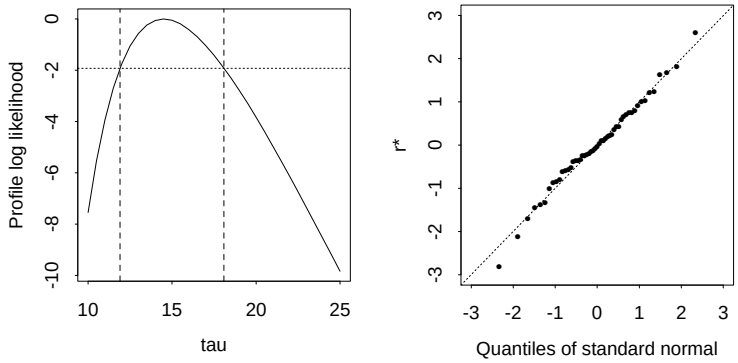
Venice data

Example 20 (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 1: 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.

2.3 Generalized Linear Models

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

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \quad y \in \mathcal{Y}, \theta \in \Omega_\theta, \phi > 0, \tag{14}$$

  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 21 (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 22 (Poisson distribution) *Write the Poisson mass function as a GLM density, and find its canonical link function.*

Example 23 (Normal distribution) *Write the normal density function as a GLM density, and find its canonical link function.*

Note to Example 21

- 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 22

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 (4) 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 23

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.

Estimation of β

Example 24 (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 24

- 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 24, 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

$$\begin{aligned} u_j &= \frac{\partial\theta_j}{\partial\eta_j}\frac{\partial\ell_j(\theta_j)}{\partial\theta_j} = \frac{y_j-\mu_j}{g'(\mu_j)\phi_j V(\mu_j)}, \\ w_j &= \left(\frac{\partial\theta_j}{\partial\eta_j}\right)^2 \frac{\partial^2\ell_j(\theta_j)}{\partial\theta_j^2} = \frac{1}{g'(\mu_j)^2\phi_j V(\mu_j)}, \end{aligned} \tag{15}$$

  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 (**??**).

- 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**

$$P = \frac{1}{\phi}\sum_{j=1}^{n}\frac{(y_j-\widehat{\mu}_j)^2}{a_j V(\widehat{\mu}_j)} \mathrel{\dot\sim} \chi^2_{n-p},$$

  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

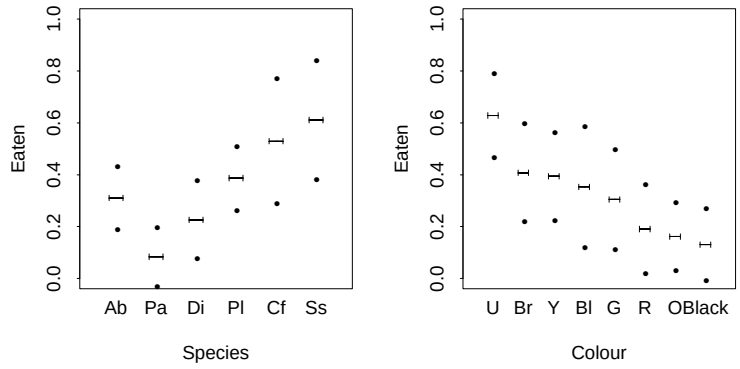
Table 3: Response (N=not sampled, S = sampled and rejected, E = eaten) of a rufous-tailed jacamar to individuals of seven species of palatable butterflies with artifically coloured wing undersides. Data from Peng Chai, University of Texas.

	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 2: 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

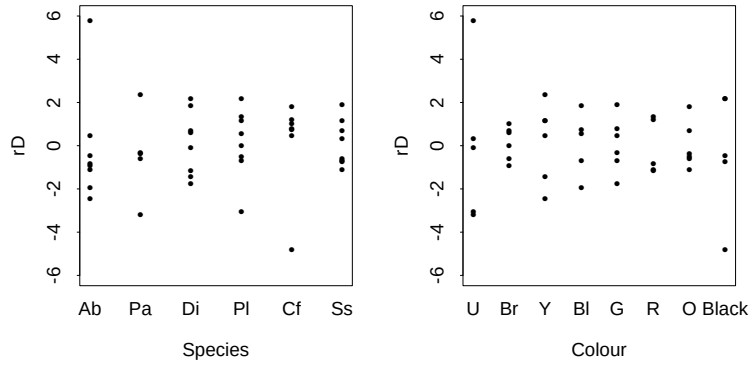
Table 4: Deviances and analysis of deviance for models fitted to jacamar data. The lower part shows results for the reduced data, without two outliers.

	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 3: Standardized deviance residuals r_D for binomial two-way layout fitted to jacamar data.



Jacamar data: Parameter estimates

Table 5: 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 6: 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 7: 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

$$\frac{6.22/3}{0.423} = 4.78 \mathrel{\dot\sim} F_{3,27}$$

  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.

2.4 Proportion Data

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

$$\mathrm{P}(Y=1)=\pi=\frac{\exp(x^{\mathrm{T}}\beta)}{1+\exp(x^{\mathrm{T}}\beta)},\quad \mathrm{P}(Y=0)=1-\pi=\frac{1}{1+\exp(x^{\mathrm{T}}\beta)},$$

leading to a linear model for the **log odds** of success,

$$\log\left\{\frac{\mathrm{P}(Y=1)}{\mathrm{P}(Y=0)}\right\}=\log\left(\frac{\pi}{1-\pi}\right)=x^{\mathrm{T}}\beta,\quad \beta\in\mathbb{R}^p.$$

- 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

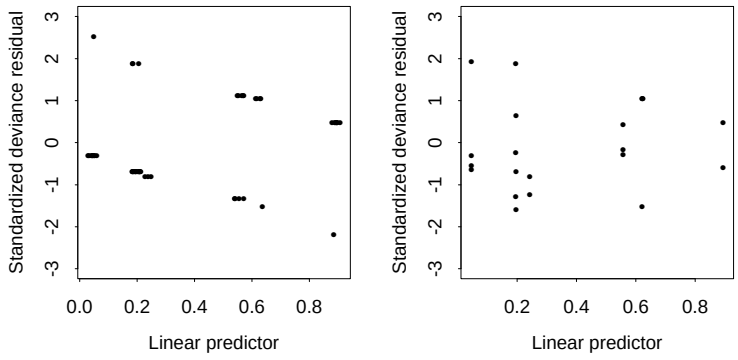
- Model with lowest AIC has stage, xray, acid:

$$x^{\mathrm{T}}\widehat{\beta} = -3.05 + 1.65 I_{\mathsf{stage}} + 1.91 I_{\mathsf{xray}} + 1.64 I_{\mathsf{acid}},$$

  where $I_{\mathsf{stage}} = 1$ 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 $I_{\mathsf{stage}} = 1$, 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 $I_{\mathsf{stage}} = I_{\mathsf{xray}} = I_{\mathsf{acid}} = 1$, 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 4: 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.

2.5 Count Data

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.

2.6 Poisson Regression

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.

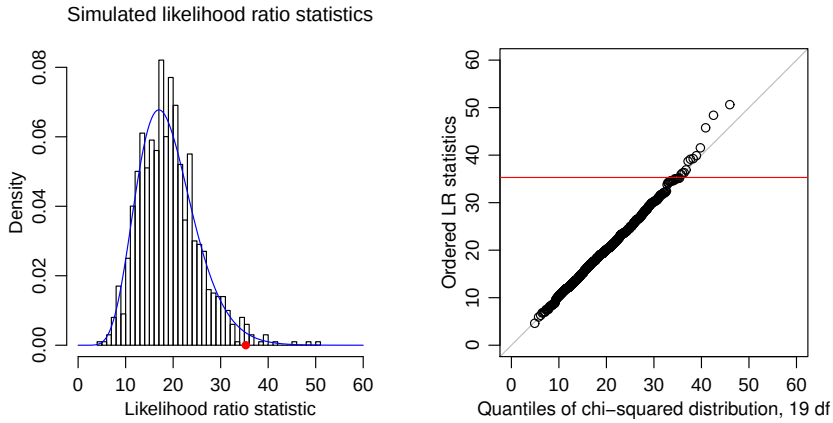
Premier League data: Analysis of deviance

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.

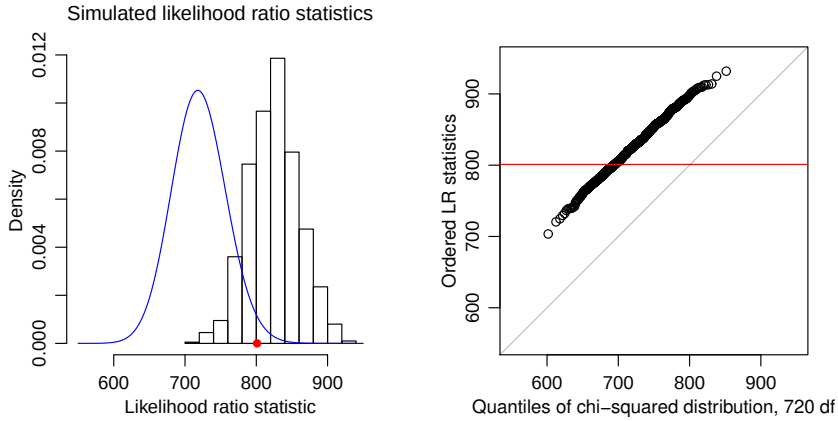
Premier League data: Null deviance for defence effect

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.



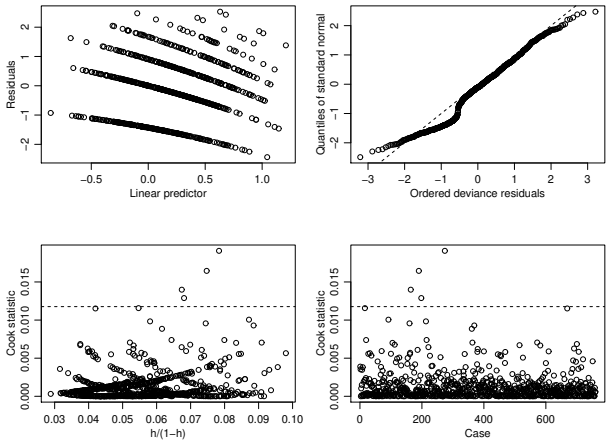
Premier League data: Estimates

	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).



2.7 Contingency Tables

Sampling schemes

- A **contingency table** contains individuals (sampling units) cross-classified by various categorical variables.
  - Example: the jacamar data cross-classify butterflies by

    $$6 \text{ species} \times 8 \text{ colours} \times 3 \text{ fates}$$

    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 25 *Find the likelihoods for each of these sampling schemes, under (unrealistic!) assumptions of independence of voters.*

Note to Example 25

- 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 26 *If parameters τ_r for the row margins are included in the above setup, then we can write*

$$\ell_{\text{Poiss}}(\beta, \tau) = \ell_{\text{Poiss}}(\tau; m) + \ell_{\text{Mult}}(\beta; y \mid m).$$

- 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 26

- The Poisson model has no conditioning, so with $\log \mu_{rc} = \gamma_r + x_{rc}^{\mathrm{T}}\beta$ the log likelihood is

$$\ell_{\text{Poiss}}(\beta, \gamma) \equiv \sum_{r,c} (y_{rc} \log \mu_{rc} - \mu_{rc}) = \sum_{r=1}^{R} \left(m_r \gamma_r + \sum_{c=1}^{C} y_{rc} x_{rc}^{\mathrm{T}}\beta - e^{\gamma_r} \sum_{c=1}^{C} e^{x_{rc}^{\mathrm{T}}\beta} \right).$$

- Now we reparametrise in terms of the row totals $\tau_r = \sum_c \mu_{rc}$, noting that

$$\tau_r = e^{\gamma_r} \sum_{c=1}^{C} e^{x_{rc}^{\mathrm{T}}\beta}, \quad \gamma_r = \log \tau_r - \log \left\{ \sum_{c=1}^{C} \exp(x_{rc}^{\mathrm{T}}\beta) \right\},$$

so

$$\begin{aligned} \ell_{\text{Poiss}}(\beta, \tau) &\equiv \sum_{r=1}^{R} (m_r \log \tau_r - \tau_r) + \sum_{r=1}^{R} \left\{ \sum_{c=1}^{C} y_{rc} x_{rc}^{\mathrm{T}}\beta - m_r \log \left(\sum_{c=1}^{C} e^{x_{rc}^{\mathrm{T}}\beta} \right) \right\}, \\ &= \ell_{\text{Poiss}}(\tau; m) + \ell_{\text{Mult}}(\beta; y \mid m), \end{aligned}$$

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!)

2.8 Ordinal Responses

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.

- ☐ We imagine that the assigned category stems from an underlying continuous variable, even if this cannot be quantified very well.

Models

- Assume we have n independent individuals whose responses $I_1, \ldots, I_n$ fall into the set $\{1, \ldots, L\}$, corresponding to L ordered categories, and that

$$\gamma_l = \mathrm{P}(I_j \leq l) = \pi_1 + \cdots + \pi_l, \quad l = 1, \ldots, L, \quad \gamma_L = 1,$$

- 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

$$I_j = l \quad \Leftrightarrow \quad x_j^{\mathrm{T}}\beta + \varepsilon_j \in (\zeta_{l-1}, \zeta_l], \quad \zeta_0 = -\infty < \zeta_1 < \cdots < \zeta_{L-1} < \zeta_L = \infty,$$

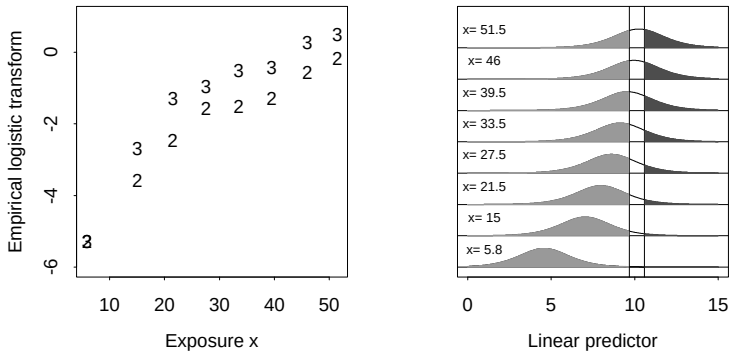
  where the tolerance distribution F of ε_j is often taken to be logistic, giving the **proportional odds model**, in which

$$\pi_l(x_j^{\mathrm{T}}\beta) = \mathrm{P}(\zeta_{l-1} < x_j^{\mathrm{T}}\beta + \varepsilon \leq \zeta_l) = F(\zeta_l - x_j^{\mathrm{T}}\beta) - F(\zeta_{l-1} - x_j^{\mathrm{T}}\beta), \quad l = 1, \ldots, L;$$

  here $\zeta_1, \ldots, \zeta_{L-1}$ are aliased with an intercept β_0 and are not usually of interest.
- Another standard tolerance distribution 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 are more interpretable statistically.
- ☐ Also possible to fit models for ordinal data, using multinomial models and tolerance distribution interpretation used for binomial data.

2.9 Overdispersion slide 166

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
Year	Quarter	$0^\dagger$	1	2	3	4	5	6	$\cdots$	$\geq$14	of 1992
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1988	1	31	80	16	9	3	2	8	$\cdots$	6	174
	2	26	99	27	9	8	11	3	$\cdots$	3	211
	3	31	95	35	13	18	4	6	$\cdots$	3	224
	4	36	77	20	26	11	3	8	$\cdots$	2	205
1989	1	32	92	32	10	12	19	12	$\cdots$	2	224
	2	15	92	14	27	22	21	12	$\cdots$	1	219
	3	34	104	29	31	18	8	6	$\cdots$		253
	4	38	101	34	18	9	15	6	$\cdots$		233
1990	1	31	124	47	24	11	15	8	$\cdots$		281
	2	32	132	36	10	9	7	6	$\cdots$		245
	3	49	107	51	17	15	8	9	$\cdots$		260
	4	44	153	41	16	11	6	5	$\cdots$		285
1991	1	41	137	29	33	7	11	6	$\cdots$		271
	2	56	124	39	14	12	7	10	$\cdots$		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)
- ☐ Chain-ladder model: number of reports in row j and column k is Poisson, with mean

$$\mu_{jk} = \exp(\alpha_j + \beta_k).$$

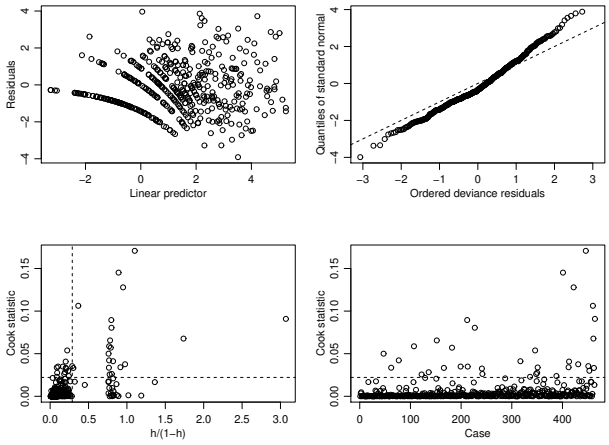
- ☐ Analysis of deviance:

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.
- ☐ Residuals on next page show no obvious problems, just generic overdispersion.

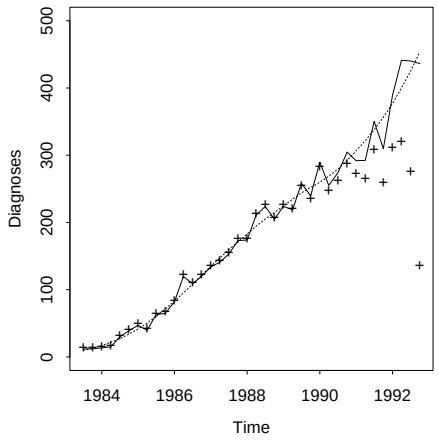
AIDS data: Assessment of fit

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 27 (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 27

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

$$\text{E}(Y) = \text{E}_\varepsilon\left\{\text{E}(Y \mid \varepsilon)\right\}, \quad \text{var}(Y) = \text{var}_\varepsilon\left\{\text{E}(Y \mid \varepsilon)\right\} + \text{E}_\varepsilon\left\{\text{var}(Y \mid \varepsilon)\right\},$$

so the response has mean and variance

$$\text{E}(Y) = \text{E}_\varepsilon(\mu\varepsilon) = \mu, \quad \text{var}(Y) = \text{var}_\varepsilon(\mu\varepsilon) + \text{E}_\varepsilon(\mu\varepsilon) = \mu(1 + \xi\mu).$$

If on the other hand the variance of ε is ξ/μ, then $\text{var}(Y) = (1 + \xi)\mu$. 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 28 (Negative binomial) *In Example 27, 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 29 (AIDS data)

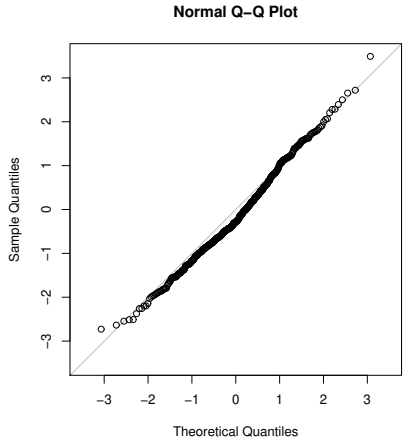
- *MLE $\widehat{\xi}_Q = 22.7$ (5.5)*
- *Analysis of Deviance (with $\widehat{\xi}_Q$ fixed):*

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 $h(\beta; Y)_{p\times 1} = 0_p$ 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 30 *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 $\{g'(\mu_j)^2\phi_j V(\mu_j)\}^{-1}$.*

- 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

$$\widehat{\phi} = \frac{1}{n-p}\sum_{j=1}^{n} \frac{(y_j - \widehat{\mu}_j)^2}{a_j g'(\mu_j)^2 V(\widehat{\mu}_j)}.$$

Note to Lemma 30

- 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 $B_j(\mu_j) = g'(\mu_j)\phi_j V(\mu_j)$. 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 jth 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 $(X^{\mathrm{T}}WX)^{-1}(X^{\mathrm{T}}W'X)(X^{\mathrm{T}}WX)^{-1}$, 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.

3 Regularisation slide 180

3.1 Basic Notions slide 181

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 $\text{rank}(X) = \min(n, p) = n$.
- Even tall X may be 'almost singular', making β 'almost inestimable'.
- Solutions:
  - subset selection (drop certain columns of X);
  - seek different good explanations of response variation, not single model;
  - regularisation (often with prediction in mind).
- Certain regularisation methods (e.g., lasso) also perform subset selection.

Different good explanations

- With $p > n$, perhaps $p \gg n$, X is rank-deficient and many β may 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 $\{1, \ldots, k\}^a$;
  - 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 most 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)

Collinearity

- ☐ Columns of X **collinear** if there exists a non-zero $v_{p\times 1}$ such that $Xv = 0$, i.e., rank$(X) < p$, so there is no unique $\widehat{\beta}$ minimising $\|y - X\beta\|^2$.
- ☐ Software deals with this by dropping columns of X, but it may be better to write $X\beta = XC\gamma$, where XC is full rank and γ has a clear interpretation.
- ☐ If X is nearly collinear, its SVD $U_{n\times n}D_{n\times p}V^{\mathrm{T}}_{p\times p}$, with $d_1 \geq \cdots \geq d_p \geq 0$, gives

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = VD_-^{\mathrm{T}}U^{\mathrm{T}}y = \sum_{r=1}^{p}(u_r^{\mathrm{T}}y/d_r)v_r,$$

  so $\widehat{\beta}$ is a linear combination of the vectors v_r with coefficients $u_r^{\mathrm{T}}y/d_r$. As $\mathrm{var}(U^{\mathrm{T}}y) = \sigma^2 I_n$,

$$\mathrm{var}(\widehat{\beta}) = \sigma^2 VD_-^{\mathrm{T}}D_-V^{\mathrm{T}} = \sigma^2\sum_{r=1}^{p} d_r^{-2}v_r v_r^{\mathrm{T}},$$

  i.e., $\widehat{\beta}$ is unstable in the directions corresponding to the v_r with small singular values d_r.
- ☐ In numerical analysis, collinearity often measured using **condition number** $(d_1/d_p)^{1/2}$, but its statistical meaning is unclear.

Regularisation

- ☐ Stop $\widehat{\beta}$ from fluctuating too wildly in directions with small eigenvalues d_r, by adding a non-negative penalty $p_\lambda(\beta)$ and choosing β to minimise the **penalised sum of squares**

$$\|y - X\beta\|^2 + p_\lambda(\beta). \tag{16}$$

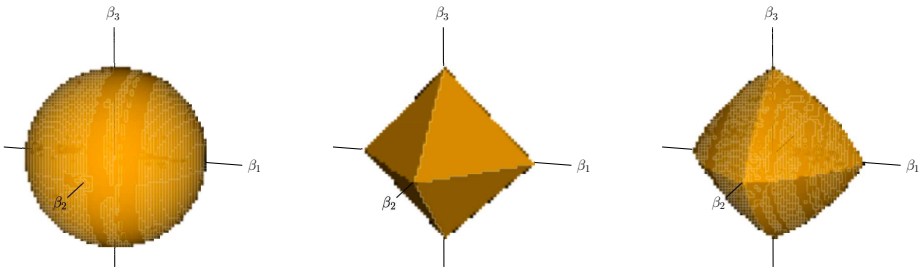
- ☐ The strength of the penalty depends on a positive parameter λ that constrains β more as λ increases.
- ☐ Often $p_\lambda(\beta) = \lambda p(\beta)$, where, for example,
  - $p(\beta) = \|\beta\|_2^2 = \sum_{r=1}^{p}\beta_r^2$ gives **ridge regression** (aka Tikhonov regularisation);
  - $p(\beta) = \|\beta\|_1 = \sum_{r=1}^{p}|\beta_r|$ gives the **lasso** (aka L_1 regularisation);
  - $p(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$ for $0 \leq \alpha \leq 1$ 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.
- ☐ It is useful to see regularisation through the lens of Bayesian inference, with the regularising term equivalent to the prior density.

Bound form

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

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

  for some $t \geq 0$, 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.



Bayesian setting

- Treat all unknowns as random variables, and compute conditional distribution of unobserved unknowns conditional on observed unknowns.
- Requires prior density on β, and if σ^2 is known, then a simple combination of **data model** and **prior model** is

$$y \mid \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n), \quad \beta \mid \sigma^2 \sim \mathcal{N}_p(\beta_*, \sigma^2 V_*), \tag{17}$$

  where the prior model is determined by β_* and V_*.
- Full specification would require prior on σ^2, but we don't need this.
- Let $\equiv$ mean we have dropped additive constants not involving the argument of a density.
- The log multivariate normal density is

$$\begin{aligned} \log f(x \mid \mu, \Omega) &= -\frac{m}{2}\log 2\pi - \frac{1}{2}\log|\Omega| - \frac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu) \\ &\equiv x^{\mathrm{T}}\Omega^{-1}\mu - \frac{1}{2}x^{\mathrm{T}}\Omega^{-1}x \\ &\equiv Q(x) = x^{\mathrm{T}}a - \frac{1}{2}x^{\mathrm{T}}Bx, \end{aligned}$$

  say, and as $\exp Q(x)$ is proportional to a unique probability density function,

$$\mathrm{E}(X) = \mu = B^{-1}a, \quad \mathrm{var}(X) = \Omega = B^{-1}, \quad \text{where } B \text{ is the } \textbf{precision matrix}.$$

Bayesian linear model I

- The model (6) gives

$$\begin{aligned}\log f(\beta \mid y, \sigma^2) &= \log\left\{\frac{f(y \mid \beta, \sigma^2) f(\beta \mid \sigma^2)}{f(y \mid \sigma^2)}\right\} \\ &\equiv \log f(y \mid \beta, \sigma^2) + \log f(\beta \mid \sigma^2) \\ &\equiv -\frac{(y - X\beta)^{\mathrm{T}}(y - X\beta)}{2\sigma^2} - \frac{(\beta - \beta_*)^{\mathrm{T}} V_*^{-1}(\beta - \beta_*)}{2\sigma^2} \\ &\propto \|y - X\beta\|_2^2 + (\beta - \beta_*)^{\mathrm{T}} V_*^{-1}(\beta - \beta_*).\end{aligned}$$

- Comparison with (5) shows that $p_\lambda(\beta)$ represents prior beliefs about the likely values of β: before seeing the data, the most plausible value is β_*, with precision V_*^{-1}.
- Dropping more constants,

$$\begin{aligned}\log f(\beta \mid y, \sigma^2) &\equiv \frac{1}{\sigma^2}\left\{\beta^{\mathrm{T}} X^{\mathrm{T}} y - \beta^{\mathrm{T}}(X^{\mathrm{T}} X)\beta/2 + \beta^{\mathrm{T}} V_*^{-1}\beta_* - \beta^{\mathrm{T}} V_*^{-1}\beta/2\right\} \\ &= \frac{1}{2\sigma^2}\left\{2\beta^{\mathrm{T}}(X^{\mathrm{T}} y + V_*^{-1}\beta_*) - \beta^{\mathrm{T}}(X^{\mathrm{T}} X + V_*^{-1})\beta\right\}, \qquad (18)\end{aligned}$$

  which is $Q(x)$ with x, a and B replaced by β, $(X^{\mathrm{T}} y + V_*^{-1}\beta_*)/\sigma^2$ and $(X^{\mathrm{T}} X + V_*^{-1})/\sigma^2$.
- Hence $f(\beta \mid y, \sigma^2)$ is multivariate normal with mean vector and variance matrix

$$\mathrm{E}(\beta \mid y, \sigma^2) = (X^{\mathrm{T}} X + V_*^{-1})^{-1}(X^{\mathrm{T}} y + V_*^{-1}\beta_*), \quad \mathrm{var}(\beta \mid y, \sigma^2) = \sigma^2 (X^{\mathrm{T}} X + V_*^{-1})^{-1}.$$

Bayesian linear model II

- The **maximum a posteriori (MAP) estimator** of β is $\mathrm{E}(\beta \mid y, \sigma^2)$, and the MAP estimator of $A_{q\times p}\beta$ is $A\mathrm{E}(\beta \mid y, \sigma^2)$, which has a posterior normal density.
- When $X^{\mathrm{T}} X$ is invertible,

$$\tilde{\beta} = \mathrm{E}(\beta \mid y, \sigma^2) = (X^{\mathrm{T}} X + V_*^{-1})^{-1}(X^{\mathrm{T}} X\widehat{\beta} + V_*^{-1}\beta_*)$$

  is an average of $\widehat{\beta}$ and β_*, weighted by $X^{\mathrm{T}} X$ and V_*^{-1}.
- The posterior precision matrix

$$\mathrm{var}(\beta \mid y, \sigma^2)^{-1} = X^{\mathrm{T}} X/\sigma^2 + V_*^{-1}/\sigma^2$$

  adds the Fisher information and the prior precision matrix, V_*^{-1}/σ^2.
- High precision corresponds to small variance, and conversely:
  - letting $V_*^{-1} \to 0$ yields an improper prior density; and
  - for large V_*^{-1} the posterior precision is essentially determined by the prior precision.

  Thus the prior density regularises $\widehat{\beta}$ by including β_* and V_*.

Improper prior density

- We only need V_* to add information in directions corresponding to small singular values of X, so we might use an **improper prior** in which V_* is singular:

$$f(\beta \mid \sigma^2) = \frac{1}{(2\pi)^{p/2}|V_*|_+^{1/2}} \exp\left\{-(\beta-\beta_*)^{\mathrm{T}} V_*^- (\beta-\beta_*)/(2\sigma^2)\right\}, \tag{19}$$

  where V_* has spectral decomposition ED_*E^{T},
  - $|V_*|_+$ denotes the product of the non-zero elements of D_*, and
  - $V_*^- = \sum_{r:d_{*r}>0} e_r e_r^{\mathrm{T}}/d_{*r}$ is a generalized inverse of V_*.
- Below we write V_*^- even when V_* is invertible.
- (8) is improper because it is not integrable in the directions of the columns of E for which the corresponding d_r^* equal zero, but we need only that the posterior density of β be proper, i.e., that the posterior precision matrix

$$\mathrm{var}(\beta \mid y, \sigma^2)^{-1} = X^{\mathrm{T}}X/\sigma^2 + V_*^-/\sigma^2$$

  is invertible.

Empirical Bayes

- Use the data to estimate the prior: construct estimators using Bayesian arguments, but assess their properties using classical criteria (bias, MSE, ...)
- The estimator $\tilde{\beta} = \mathrm{E}(\beta \mid y, \sigma^2)$ has mean and variance

$$\begin{aligned} \mathrm{E}(\tilde{\beta} \mid \beta) &= (X^{\mathrm{T}}X + V_*^-)^{-1}(X^{\mathrm{T}}X\beta + V_*^-\beta_*) \\ &= \beta + (X^{\mathrm{T}}X + V_*^-)^{-1}V_*^-(\beta_* - \beta), \\ \mathrm{var}(\tilde{\beta} \mid \beta) &= \sigma^2(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X + V_*^-)^{-1}. \end{aligned} \tag{20}$$

- Hence $\tilde{\beta}$
  - is biased unless $\beta_* = \beta$,
  - has smaller variance than $\widehat{\beta}$,

  so maybe there is a bias-variance tradeoff when estimating $A\beta$.
- If we write $\mu = \mathrm{E}(\tilde{\beta} \mid \beta)$, then the MSE is

$$\begin{aligned} \mathrm{E}\left(\|A\tilde{\beta} - A\beta\|^2 \mid \beta\right) &= \mathrm{E}\{(\tilde{\beta}-\beta)^{\mathrm{T}}A^{\mathrm{T}}A(\tilde{\beta}-\beta) \mid \beta\} \\ &= \mathrm{E}\left[\mathrm{tr}\left\{A(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{\mathrm{T}}A^{\mathrm{T}}\right\} \mid \beta\right] \\ &= \mathrm{tr}\left[\mathrm{E}\left\{A(\tilde{\beta}-\mu+\mu-\beta)(\tilde{\beta}-\mu+\mu-\beta)^{\mathrm{T}}A^{\mathrm{T}} \mid \beta\right\}\right]. \end{aligned}$$

Empirical Bayes II

- The expectation above is

$$A\left\{\mathrm{var}(\tilde{\beta} \mid \beta) + (X^{\mathrm{T}}X + V_*^-)^{-1}V_*^-(\beta - \beta_*)(\beta - \beta_*)^{\mathrm{T}}V_*^-(X^{\mathrm{T}}X + V_*^-)^{-1}\right\}A^{\mathrm{T}},$$

  giving the MSE when estimating a fixed β.
- Taking expectations over the prior model for β gives

$$\mathrm{E}\left(\|A\tilde{\beta} - A\beta\|^2\right) = \sigma^2 \mathrm{tr}\left\{A(X^{\mathrm{T}}X + V_*^-)^{-1}A^{\mathrm{T}}\right\}, \tag{21}$$

  which is larger than $A\mathrm{var}(\tilde{\beta} \mid \beta)A^{\mathrm{T}}$ and does not depend on β_*.
- This computation uses only the mean and variance, so holds under second-order assumptions.
- From now on we set $\beta_* = 0$, unless we state otherwise.

Equivalent degrees of freedom

- If we set $\beta_* = 0$, then the fitted values are

$$\tilde{y} = X\tilde{\beta} = X(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}y = H_*y,$$

  say.
- We define the **equivalent degrees of freedom** of the fit as

$$\mathrm{edf} = \mathrm{tr}(H_*) = \mathrm{tr}\{X(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}\} = p - \mathrm{tr}\{(X^{\mathrm{T}}X + V_*^-)^{-1}V_*^-\},$$

- This is lower than p unless $V_*^- = 0$, so regularisation reduces the degrees of freedom by an amount that depends on V_*.
- The penalised estimate is a linear function of the unpenalised one (if it exists), as we can write

$$\tilde{\beta} = (X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}X\widehat{\beta} = P_*\widehat{\beta},$$

  say. As

$$\mathrm{edf} = \mathrm{tr}(H_*) = \mathrm{tr}(P_*),$$

  this gives an alternative formula useful in complex models.

How much penalisation?

- ☐ Often V_*^- depends on some $\lambda > 0$ that must be chosen, as well as σ^2, which is usually estimated by a (penalised) residual sum of squares.
- ☐ To estimate λ, we compare y_j with its predicted value $\widehat{y}_{\lambda,j}^- = x_j^{\mathrm{T}}\widehat{\beta}_{\lambda,-j}$, where $\widehat{\beta}_{\lambda,-j}$ is

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

  computed with the jth rows x_j and y_j of X and y omitted.
- ☐ Using Lemma 14, the **leave-one-out cross-validation** sum of squares is then

$$\mathrm{CV}_\lambda = \sum_{j=1}^n (y_j - \widehat{y}_{\lambda,j}^-)^2 = \|y - \widehat{y}_\lambda^-\|^2 = \sum_{j=1}^n \frac{(y_j - \widehat{y}_{\lambda,j})^2}{(1-h_{\lambda,jj})^2},$$

  where $\widehat{y}_{\lambda,j}$ is the jth element of the complete-data fitted value $H_\lambda y$ and $h_{\lambda,jj}$ is the jth diagonal element of $H_\lambda = X(X^{\mathrm{T}}X + V_*^-)^{-1}X^{\mathrm{T}}$ for the overall fit.
- ☐ More often we use the **generalized cross-validation** criterion

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

- ☐ Whichever criterion is used, it is typically minimised numerically over a grid of values of λ.

REML

- ☐ Cross-validation makes only second-order assumptions.
- ☐ Under normality, the marginal density of y is $\mathcal{N}\{X\beta_*, \sigma^2(I_n + XV_*X^{\mathrm{T}})\}$, so we could estimate β_*, σ^2 and λ by maximising the corresponding likelihood.
- ☐ If n and p are large, this results in biased estimates of λ and σ^2, so we prefer to eliminate β_*, resulting in a **log restricted likelihood** whose form is given below, with $W_\lambda^{-1} = I_n + XV_*X^{\mathrm{T}}$.

Lemma 31 *In a model in which $y \sim \mathcal{N}(X\beta, \sigma^2 W_\lambda^{-1})$, where W_λ depends on a parameter λ, a log restricted likelihood for σ^2 and λ is*

$$\ell_{\mathrm{REML}}(\sigma^2, \lambda) \equiv \frac{1}{2}\log(|W_\lambda|/|X^{\mathrm{T}}W_\lambda X|) - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - \widehat{y}_\lambda)^{\mathrm{T}}W_\lambda(y - \widehat{y}_\lambda),$$

where $\widehat{\beta}_\lambda = (X^{\mathrm{T}}W_\lambda X)^{-1}X^{\mathrm{T}}W_\lambda y$ and $\widehat{y}_\lambda = X\widehat{\beta}_\lambda$. For fixed λ the restricted maximum likelihood estimator of σ^2 is therefore

$$\widehat{\sigma}_\lambda^2 = \frac{1}{n-p}(y - \widehat{y}_\lambda)^{\mathrm{T}}W_\lambda(y - \widehat{y}_\lambda),$$

and the resulting profile log restricted likelihood for λ is

$$\ell_{\mathrm{p}}(\lambda) \equiv \frac{1}{2}\log(|W_\lambda|/|X^{\mathrm{T}}W_\lambda X|) - \frac{(n-p)}{2}\log\widehat{\sigma}_\lambda^2.$$

Note on Lemma 31

- Suppose that $f(y;\alpha,\beta)$ depends on two parameters, that interest is focused on α, and that for fixed α there is a minimal 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 use it for inference on α, in the form

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

  As the left-hand side of this expression does not depend on β, we may be able to simplify the right-hand side by an astute choice of β.

- In the normal model we take $\alpha = (\sigma^2,\lambda)$. If α is fixed, then $s_\alpha = \widehat{\beta}_\alpha = (X^{\mathrm{T}} W_\lambda X)^{-1} X^{\mathrm{T}} W_\lambda y$ is sufficient for β; its distribution is $\mathcal{N}_p\{\beta, \sigma^2 (X^{\mathrm{T}} W_\lambda X)^{-1}\}$. Hence

$$\ell_{\mathrm{REML}}(\sigma^2,\lambda) = \log f(y \mid \widehat{\beta}_\lambda;\sigma^2,\lambda) = \log f(y;\sigma^2,\lambda,\beta) - \log f(\widehat{\beta}_\lambda;\sigma^2,\lambda,\beta)$$

  which equals

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

  or equivalently, on setting $\beta = 0$ and $\widehat{y}_\lambda = X\widehat{\beta}_\lambda$,

$$\tfrac{1}{2}\log(|W_\lambda|/|X^{\mathrm{T}} W_\lambda X|) - \frac{(n-p)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y^{\mathrm{T}} W_\lambda y - \widehat{y}_\lambda^{\mathrm{T}} X^{\mathrm{T}} W_\lambda \widehat{y}_\lambda\right).$$

- The last term reduces to the given form because $\widehat{y}_\lambda^{\mathrm{T}} W_\lambda (y - \widehat{y}_\lambda) = 0$, so the term in brackets in the last displayed equation is the residual sum of squares $(y - \widehat{y}_\lambda)^{\mathrm{T}} W_\lambda (y - \widehat{y}_\lambda)$.

- The restricted maximum likelihood estimator $\widehat{\sigma}^2_\lambda$ and the profile log restricted likelihood for λ are obtained by maximising $\ell_{\mathrm{REML}}(\sigma^2,\lambda)$, for fixed λ and then dropping constant terms from $\ell_{\mathrm{REML}}(\widehat{\sigma}^2_\lambda,\lambda)$.

Numerical example from Wood (2011, JRSSB)

The usual methods (AIC, GCV, ...) for choosing λ are available, but we focus on likelihood methods; see below.

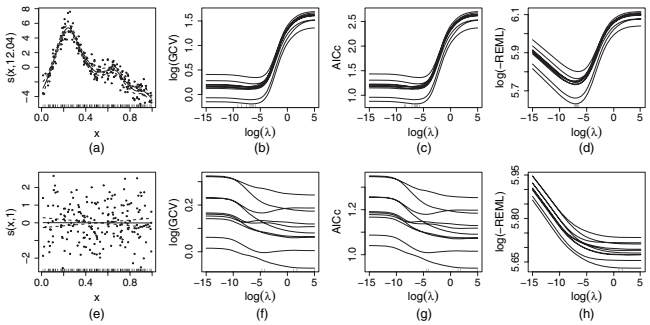


Fig. 1. Example comparison of GCV, AICc and REML criteria: (a) some (x,y)-data modelled as $y_i = f(x_i) + \varepsilon_i$, ε_i independent and identically distributed $N(0,\sigma^2)$ where smooth function f was represented by using a rank 20 thin plate regression spline (Wood, 2003); (b)–(d) various 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)

3.2 Simple Applications

Ridge regression

- ☐ Used for prediction when X is close to singular.
- ☐ If the first column of X is 1_n, we set $\beta_* = 0$ and $V_*^- = \lambda S = \lambda \mathrm{diag}(0, I_{p-1})$, giving

$$\widehat{\beta}_\lambda = (X^{\mathrm{T}} + \lambda S)^{-1} X^{\mathrm{T}} y, \quad \widehat{y}_\lambda = X\widehat{\beta}_\lambda = X(X^{\mathrm{T}} + \lambda S)^{-1} X^{\mathrm{T}} y = H_\lambda y,$$

  and effective degrees of freedom

$$\mathrm{edf}_\lambda = \mathrm{tr}(H_\lambda) = \mathrm{tr}\{(X^{\mathrm{T}}X + \lambda S)^{-1} X^{\mathrm{T}} X\} = \sum_{r=1}^{p} \frac{1}{1 + \lambda\delta_r},$$

  where $\delta_p \geq \cdots \geq \delta_2 > \delta_1 = 0$ are the eigenvalues of $(X^{\mathrm{T}}X)^{-1/2} S (X^{\mathrm{T}}X)^{-1/2}$.
- ☐ As λ increases from zero to infinity, edf_λ decreases from $p = \mathrm{rank}(X)$ to 1. The two are equivalent, but edf_λ is more easily interpreted, because it is not related to the scale of X.
- ☐ The inverse exists even if $X^{\mathrm{T}}X$ is singular, but if it is invertible then

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

  so as $\lambda \to \infty$ all the elements of $\widehat{\beta}_\lambda$ tend to zero, other than the first. This corresponds to reducing the prior variance to zero, thereby giving the data themselves less and less influence on the elements of $\widehat{\beta}_\lambda$ other than the first, and thus stabilises the estimator.

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
```

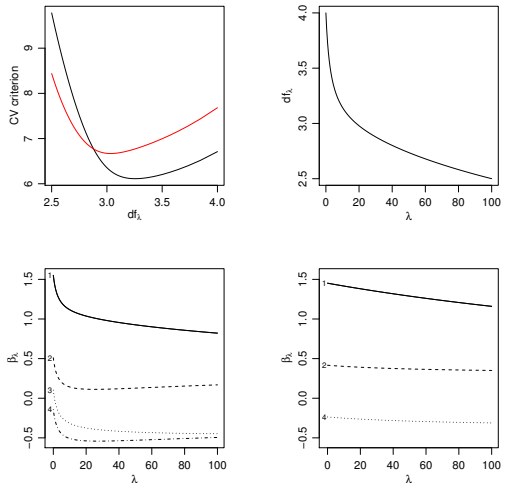

Example: Cement data

	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.
- Software often preprocesses X and y by either
  - centering both, by subtracting column means, or
  - centering y and centering and scaling X, so the column means are zero and the column variances are unity.
- 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

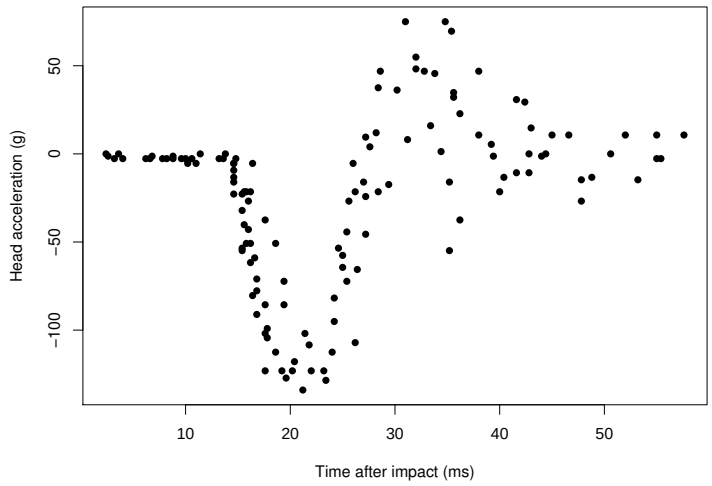
- ☐ The literature on ridge regression is very large and very dispersed, with many variants and many connections to ML techniques.
- ☐ Be careful with software: any pre-processing of X is not always described.

Semiparametric regression

- ☐ 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?'
- ☐ Most basic tool is the **scatterplot smoother**.

Example: Motorcycle data

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



Scatterplot smoothing

- Have data $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_- \leq x_1 < \cdots < x_n \leq x_+$ (ahem) and we wish to estimate $\mathrm{E}(y) = \mu(x)$, for $x \in \mathcal{X} = [x_-, x_+]$.
- Suppose that $\mu \in \mathcal{M}$, a function space spanned by n linearly independent basis functions that can be identified by evaluation at $x_1, \ldots, x_n$, and let $\mu_j = \mu(x_j)$.
- Can choose a basis $\{b_1(x), \ldots, b_n(x)\}$ for $\mathcal{M}$ such that $\mu(x) = \sum_{j=1}^n \mu_j b_j(x)$ interpolates $(x_1, \mu_1), \ldots, (x_n, \mu_n)$.
- Suppose that $\mathcal{M}$ contains the linear functions on $\mathcal{X}$ and that the second derivatives of the $b_j(x)$ are not all zero, so functions in $\mathcal{M}$ may also be nonlinear in x.
- To estimate μ we minimise a **penalised sum of squares**,

$$\sum_{j=1}^{n} \{y_j - \mu(x_j)\}^2 + \lambda \int_{\mathcal{X}} \{\mu''(x)\}^2 \, \mathrm{d}x, \tag{22}$$

  where the **roughness penalty** imposes smoothness: if $\lambda \to 0$, then $\mu(x_j) \to y_j$ and $\widehat{\mu}$ interpolates, but when $\lambda \to \infty$ even tiny wiggles in μ will give a huge penalty, making $\widehat{\mu}$ linear.
- The penalty does not affect linear functions, so $\mathcal{M} = \mathcal{L} \bigoplus \mathcal{P}$, where $\mathcal{L}$ and $\mathcal{P}$ are the two-dimensional vector space of linear functions on $\mathcal{X}$ and an $(n-2)$-dimensional vector space of nonlinear functions on $\mathcal{X}$, and $\bigoplus$ denotes addition of vector spaces.

Scatterplot smoothing II

- The roughness term is

$$\int_{\mathcal{X}} \{\mu''(x)\}^2 \, \mathrm{d}x = \int_{\mathcal{X}} \left\{ \sum_{j=1}^{n} \mu_j b_j''(x) \right\}^2 \mathrm{d}x = \sum_{i,j=1}^{n} \mu_i \mu_j \int_{\mathcal{X}} b_i''(x) b_j''(x) \, \mathrm{d}x = \mu^{\mathrm{T}} S \mu,$$

  say, where $\mu^{\mathrm{T}} = (\mu_1, \ldots, \mu_n)$.
- $S_{n \times n}$ has (i, j) element $\int_{\mathcal{X}} b_i''(x) b_j''(x) \, \mathrm{d}x$ and is symmetric and positive semi-definite of rank $n - 2$, because linear functions are unpenalised, so $S 1_n = S(x_1, \ldots, x_n)^{\mathrm{T}} = 0$.
- The penalised sum of squares

$$(y - \mu)^{\mathrm{T}}(y - \mu) + \lambda \mu^{\mathrm{T}} S \mu \equiv -2\mu^{\mathrm{T}} y + \mu^{\mathrm{T}}(I_n + \lambda S)\mu,$$

  is minimised by $\widehat{\mu}_\lambda = (I_n + \lambda S)^{-1} y$.
- As λ increases from zero, the fitted value $\widehat{\mu}_\lambda$ shrinks from y towards the straight-line regression fit to y, which is unpenalised.
- The equivalent degrees of freedom are $\mathrm{edf}_\lambda = \mathrm{tr}(H_\lambda) = \sum_{j=1}^{n} (1 + \lambda \delta_j)^{-1}$, where $\delta_1 \geq \cdots \geq \delta_3 > \delta_2 = \delta_1 = 0$ are the eigenvalues of S. As λ increases edf_λ decreases monotonically from $\mathrm{edf}_0 = n$ towards $\mathrm{edf}_\infty = 2$.

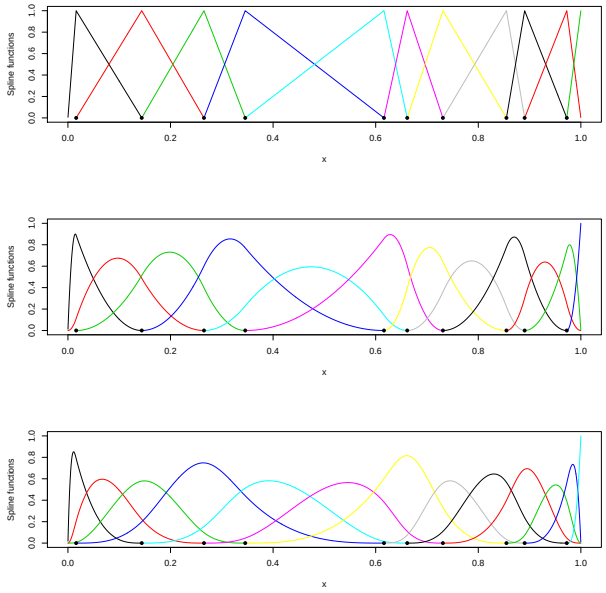
Scatterplot smoothing III

- In principle we might take any basis functions, but in practice we usually take local polynomials known as **splines** that have good approximation properties.
- There are many forms of splines, which
  - are often cubic polynomials with finite support between values of x known as **knots**, $x_1^*, \ldots, x_K^*$, and then S is tri-diagonal,
  - sometimes form a **natural cubic spline**, which has $K = n$ and certain optimality properties,
  - are discussed in more detail later.
- If there is no penalisation ($\lambda = 0$) then we have a standard linear model, and spline basis functions are called **regression splines**.
- Under second-order assumptions we choose λ by minimising $\mathrm{CV}(\lambda)$ or $\mathrm{GCV}(\lambda)$.
- Under normal-theory assumptions we can use REML to estimate σ^2 and λ.
- 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

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

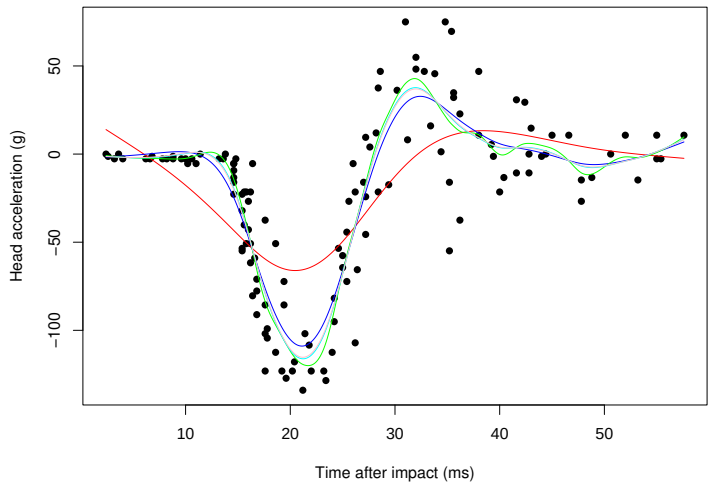
  where $S_{n' \times n'}$ arises as before from the roughness penalty on $\mu(x)$.

Linear, quadratic and cubic B-splines



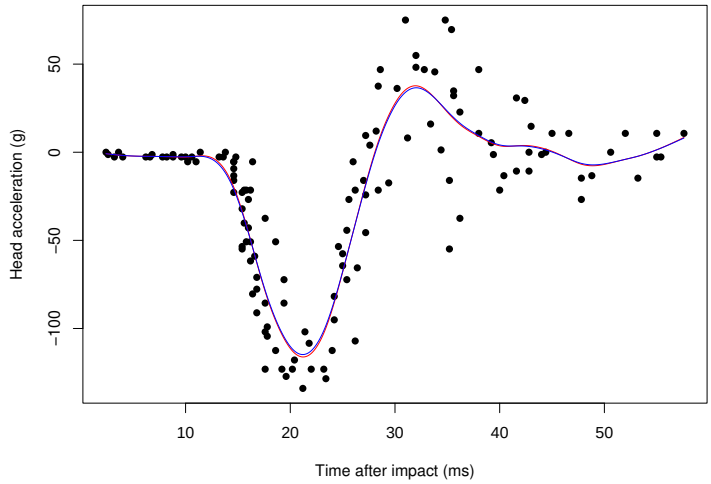
Example: Motorcycle data

Scatterplot smooths based on natural cubic splines with edf equal to 5 (red), 10 (blue), 20 (green), and chosen by CV (cyan, $\mathrm{edf} = 12.8$) and GCV (pink, $\mathrm{edf} = 12.26$):



Example: Motorcycle data

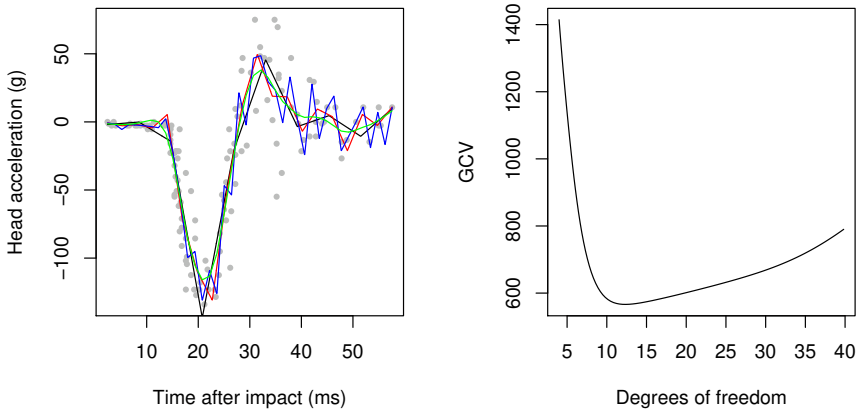
Scatterplot smooths based on natural cubic splines with weights 16 when $x \leq 12$ and 1 for $x > 12$, and edf chosen by CV (red, $\mathrm{edf} = 14.7$) and GCV (blue, $\mathrm{edf} = 13.7$):



Choosing K and λ

- ☐ Above we took $K = n$ basis functions, but for statistical purposes we seek a summary of the data, so we hope that $\mathrm{edf} \ll n$, so we hope that $K < n$, maybe even $K \ll n$.
- ☐ Theory suggests that as $n \to \infty$ we need $K = O(n^{1/5})$ or even $O(n^{1/9})$ to get near-optimal estimation of $\mu(x)$, when μ lies in reasonable function classes;
- ☐ In practice we take K (more than) large enough to give enough flexibility (increasing it if results are suspect, $K = 9$ by default in `mgcv`), and allow λ to determine the smoothness of the curve;
- ☐ Typically the knots x_k^* are placed at equally-spaced quantiles of x.

Example: Motorcycle data



- Left: linear spline fits with $\lambda = 0$ and $K = 10$ (black), 20 (red), 40 (blue), and optimal GCV choice of λ with $K = 40$ (green)
- Right: $\mathrm{GCV}(\lambda)$ as a function of df_λ for $K = 40$.

Comments

- We discuss inference (beyond 'point' estimation) and adaptive estimation of weights later ...
- Here we are producing point estimates; later we discuss the construction of confidence sets.
- An alternative local averaging approach uses locally weighted fits, such as the **Nadaraya–Watson estimator**

$$\widehat{\mu}(x) = \frac{\sum_{j=1}^{n} K\{(x - x_j)/h\} y_j}{\sum_{j=1}^{n} K\{(x - x_j)/h\}},$$

  where
  - the **kernel function** K is something like the Gaussian density, and
  - the **bandwidth** h plays a role similar to edf.

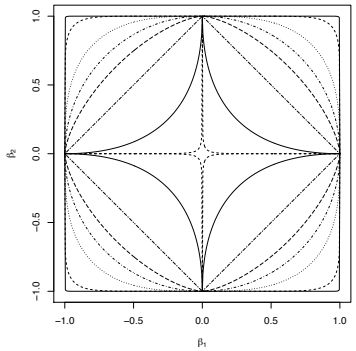
  This is also a linear smoother, and in fact the spline smoothers have representations in terms of equivalent kernels.
- Local averaging can be extended to **local likelihood** fitting of more complex models.

L_q penalties

☐ The quadratic penalty $\|\beta\|_2$ generalises to other L_q penalties

$$\|\beta\|_q = \sum_{r=1}^{p} |\beta_r|^q,$$

shown below for $p = 2$ and (working inwards) $q = 100$, 10, 3, 2, 1.5, 1, 0.5, 0.2; $\|\beta\|_0 = \#\{\beta_r \neq 0\}$ counts the number of non-zero parameters.



(Some picture credits here and later: Simon Wood)

Basic geometry

☐ If $D(\beta)$ is a sum of squares or negative log likelihood, then

$$\tilde{\beta}_\lambda = \text{argmin}_\beta \{D(\beta) + \lambda\|\beta\|_q\},$$

- satisfies $\|\tilde{\beta}_\lambda\|_q = t$ for some t, and
- minimises $D(\beta)$ on that contour, i.e.,

$$\tilde{\beta}_\lambda = \text{argmin}_\beta D(\beta) \quad \text{such that} \quad \|\tilde{\beta}_\lambda\|_q = t,$$

because otherwise we could reduce $D(\beta)$ while leaving the penalty unchanged, i.e., $\tilde{\beta}_\lambda$ would not be optimal.

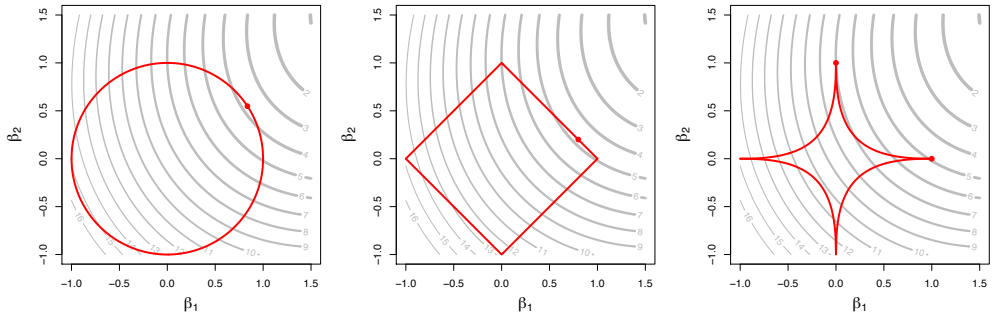
☐ The sets $\|\tilde{\beta}_\lambda\|_q = t$

- have cusps (and thus can set $\beta_r = 0$) when $q \leq 1$,
- are non-convex (and thus may give non-unique solutions) when $q < 1$,

so there is a unique solution if the contours of $D(\beta)$ and $\|\beta\|_q$ are convex, and both a unique solution and the possibility of choosing variables (sparsity) by setting $\beta_r = 0$ when $q = 1$.

Basic geometry II

Penalised solutions (red dots) for $q = 2$, 1, 0.45, with contours of $D(\beta)$ in grey and solution contour for $\|\beta\|_q$ in red.



As $\lambda \to \infty$ the constraint tightens and the red contours shrink around the origin, and as $\lambda \to 0$ the constraint relaxes and the $\tilde{\beta}_\lambda$ tends to the unconstrained estimate.

Lasso

- The **lasso (least absolute shrinkage and selection operator)** objective function can be written as

$$L = \tfrac{1}{2}\|y - X\beta\|_2 + \lambda\|\beta\|_1,$$

so suppose we have minimised this for some λ_0, giving **active set** $A = \{r : \tilde{\beta}_r \neq 0\}$ and

$$L = \tfrac{1}{2}(y - X_A\tilde{\beta}_A)^{\mathrm{T}}(y - X_A\tilde{\beta}_A) + \lambda\sum_{r\in A}|\tilde{\beta}_r|,$$

and now we aim to decrease λ (i.e., to relax the constraint).

- Now $\mathrm{d}|x|/\mathrm{d}x = \mathrm{sign}(x)$, so when

$$\frac{\mathrm{d}L}{\mathrm{d}\tilde{\beta}_A} = X_A^{\mathrm{T}}(X_A\tilde{\beta}_A - y) + \lambda\,\mathrm{sign}(\tilde{\beta}_A) = 0,$$

we have

$$\tilde{\beta}_A = (X_A^{\mathrm{T}}X_A)^{-1}X_A^{\mathrm{T}}y - \lambda(X_A^{\mathrm{T}}X_A)^{-1}\mathrm{sign}(\tilde{\beta}_A) = b - \lambda a,$$

say, i.e., $\tilde{\beta}_A$ is linear in λ until A changes.

- A changes on deleting a column X_r from X_A or on adding one from its complement X_{A^c}.
- $\mathrm{sign}(\tilde{\beta}_A)$ only changes when (say) $\tilde{\beta}_r$ passes through zero, but r leaves A when $\tilde{\beta}_r = 0$.

Lasso algorithm

- A variable in A is deleted if a component of $\tilde{\beta}_A = b - \lambda a$ hits zero as λ decreases from λ_0, which occurs at $\lambda_- = \max_{\lambda<\lambda_0} b_r/a_r$.
- If X_r is the rth column of X, then r will enter A if adding $X_r\beta_r$ decreases L, i.e., if

$$\frac{\mathrm{d}L}{\mathrm{d}\beta_r} = X_r^{\mathrm{T}}(X\beta - y) + \lambda \mathrm{sign}(\beta_r) \quad \begin{cases} < 0, & \beta_r > 0, \\ > 0, & \beta_r < 0, \end{cases}$$

  so β_r remains inactive if $|X_r^{\mathrm{T}}(y - X\beta)| \leq \lambda$.
- Thus as λ decreases, A changes when for some r in the complement A^c of A we have

$$X_r^{\mathrm{T}}(y - X_A\tilde{\beta}_A) = \pm\lambda,$$

  or, setting $\tilde{\beta}_A = b - \lambda a$,

$$X_{A^c}^{\mathrm{T}}(y - X_A b) + \lambda(X_{A^c}^{\mathrm{T}} X_A a \pm 1) = 0 \implies c + \lambda(d \pm 1) = 0,$$

  say: the next variable is added when $\lambda = \lambda_+ = \max_{\lambda<\lambda_0}\{-c_r/(d_r \pm 1)\}$.
- Hence if $s = \mathrm{sign}(\beta)$, the algorithm decreases λ from
  - the highest λ at which the a first variable is active, and defines the A and s, then
  - finds the next λ at which A changes, stores it and the corresponding $\tilde{\beta}$, updating A and s.

Practical matters and thresholding

- Usually
  - λ is chosen by dividing the data into training and testing subsets and minimising some measure of prediction error for the test subset,
  - y is centered and X has no column of ones, and
  - the columns of X are standardized to have zero mean and unit variance — what this means in terms of interpreting the components of β is then unclear!
- We can think of penalised estimators as using different sorts of **thresholding** functions, where $\widehat{\beta}$ is replaced by $\tilde{\beta} = g_\lambda(\widehat{\beta})$ and (conceptually)
  - for the **lasso** there is soft thresholding,

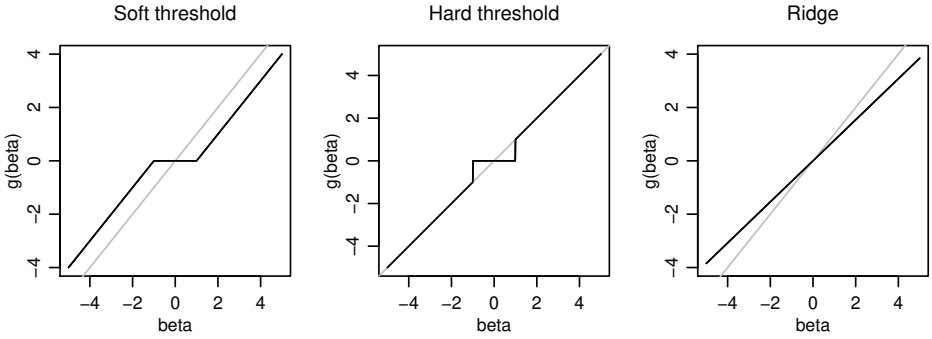
$$g_\lambda(u) = \begin{cases} 0, & |u| < \lambda, \\ \mathrm{sign}(u)(|u| - \lambda), & \text{otherwise}, \end{cases}$$

  - for **variable selection** there is hard thresholding,

$$g_\lambda(u) = \begin{cases} 0, & |u| < \lambda, \\ u, & \text{otherwise}, \end{cases}$$

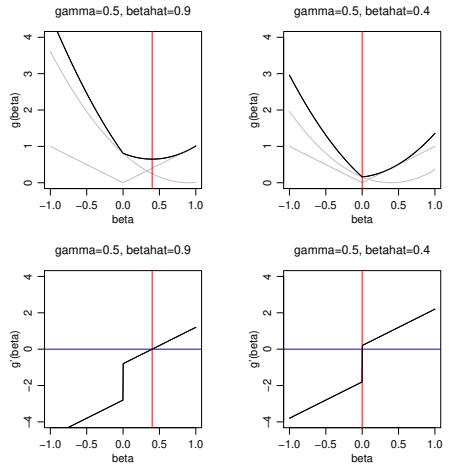
  - for **ridge regression** there is shrinkage, $g(u) = u/(1 + \lambda)$.

Threshold functions



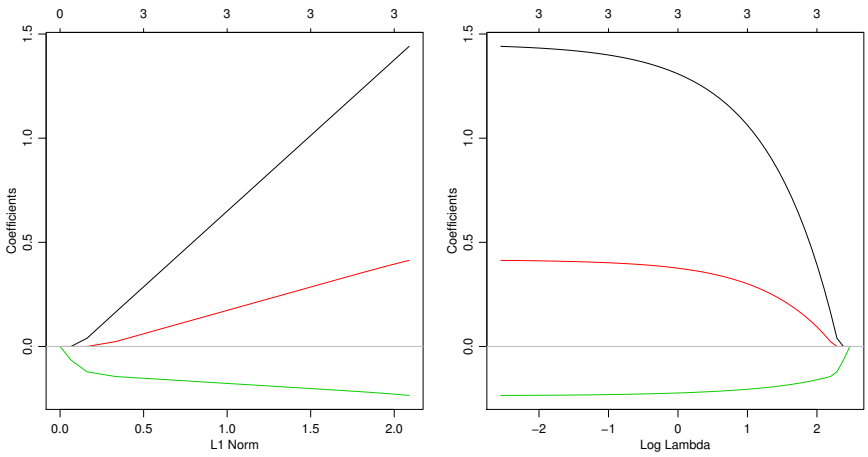
Soft thresholding

Top panels: the sum $g(\beta)$ of the L_1 penalty and the least squares function (both in grey) is the black line, which has a cusp at $\beta = 0$. If the left- and right-hand derivatives of the sum are equal at zero, then the minimiser (at the red vertical line) is non-zero, but not otherwise. Bottom panels: the derivative $g'(\beta) = 0$ when $\beta = \tilde{\beta}$.



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 can compute the lasso solution path for all λ in $O(n^3)$ operations (faster than ridge, $O(np^2)$, when $p \gg n$).
- **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 lasso variable is consistent (i.e., the probability that the variables with $\beta_r \neq 0$ are selected tends to 1), but that the $\tilde{\beta}_\lambda$ themselves 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).
- 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.
- When $D(\beta)$ is a general loss function (e.g., a negative log likelihood for a GLM), the exact algorithm above is replaced by a **coordinate descent algorithm** that updates each $\tilde{\beta}_r$ in turn, with the other components fixed. This too is very efficient.

3.4 Splines

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
$$\mu_{n\times 1} = B_{n\times p}\beta_{p\times 1}, \quad \operatorname{rank}(B) = p \leq n,$$
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.
- ☐ Focus on **spline functions**, on which there is a huge literature.

Aside: Polynomial regression

- ☐ Classical approach is to fit a 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 $\widehat{\beta}_{p\times 1} = (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}y$, where (j,i) element of $n \times p$ matrix B 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 infinite differentiability everywhere is generally unnecessarily restrictive.

Piecewise linear basis

☐ Place **knots** of a univariate x at $x_1^* < \cdots < x_K^*$, and define **tent functions**

$$
\begin{aligned}
b_1(x) &= \begin{cases} (x_2^* - x)/(x_2^* - x_1^*), & x_1^* \le x \le x_2^*, \\ 0, & \text{otherwise,} \end{cases} \\
b_k(x) &= \begin{cases} (x - x_{k-1}^*)/(x_k^* - x_{k-1}^*), & x_{k-1}^* < x \le x_k^*, \\ (x_{k+1}^* - x)/(x_{k+1}^* - x_k^*), & x_k^* < x \le x_{k+1}^*, \end{cases} \quad k = 2, \ldots, K-1, \\
b_K(x) &= \begin{cases} (x - x_{K-1}^*)/(x_K^* - x_{K-1}^*), & x_{K-1}^* \le x \le x_K^*, \\ 0, & \text{otherwise :} \end{cases}
\end{aligned}
$$

these are non-zero only in (x_{k-1}^*, x_{k+1}^*) (**compact support**) and take value 1 at x_k^*.

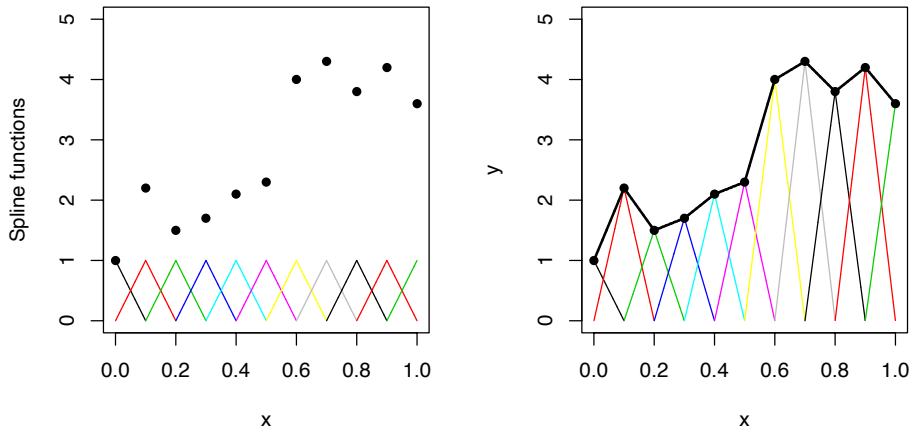
☐ An exact linear interpolant of data $y_1, \ldots, y_K$ at the knots is the function

$$\mu(x) = \sum_{k=1}^{K} b_k(x) y_k = B(x)^{\mathrm{T}} y,$$

which by construction

- passes through the points (x_k^*, y_k) and
- is linear between the knots.

Piecewise linear basis II



☐ Left: piecewise linear basis functions $b_k(x)$ and data (x_k^*, y_k).

☐ Right: functions $b_k(x)y_k$ and linear interpolant (bold).

Statistical use

- Aim for summary of the n observations, so interpolation not useful.
- Could use $K < n$ knots, but fit tends to depend heavily on their locations, so better to use high(ish) K and impose structure by penalising roughness of $\mu(x)$:

$$\widehat{\beta}_\lambda = \text{argmin}_\beta \left\{ \|y - B\beta\|^2 + \lambda \sum_{k=2}^{K-1} \left\{ \mu(x^*_{k-1}) - 2\mu(x^*_k) + \mu(x^*_{k+1}) \right\}^2 \right\}.$$

- The second term sums squared numerical second derivatives at the internal knots, and λ imposes the degree of penalisation:
  - $\lambda = 0$ (no penalty) gives the interpolant,
  - $\lambda \to \infty$ forces the second derivatives to be zero, so gives a straight-line fit.
- On setting $\beta_k = \mu(x^*_k)$ and writing

$$\begin{pmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \beta_3 - 2\beta_4 + \beta_5 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \end{pmatrix} = D_{(K-2)\times K}\beta_{K\times 1},$$

  the penalty is $\sum_{k=2}^{K-1}(\beta_{k-1} - 2\beta_k + \beta_{k+1})^2 = (D\beta)^{\mathrm{T}} D\beta = \beta^{\mathrm{T}} D^{\mathrm{T}} D\beta = \beta^{\mathrm{T}} S\beta$, say.

Penalized fit

- The penalty matrix S is of side $K \times K$ but of rank $K - 2$, because

$$S1_K = Sx^*_{K\times 1} = 0_K :$$

  the null space of S consists of all straight lines $\beta_0 1_K + \beta_1 x^*$, which are unpenalised.
- Hence (recalling ridge regression),

$$\widehat{\beta}_\lambda = \text{argmin}_\beta \left\{ \|y - B\beta\|^2 + \lambda\beta^{\mathrm{T}} S\beta \right\} = (B^{\mathrm{T}}B + \lambda S)^{-1} B^{\mathrm{T}} y$$

  giving

$$\begin{aligned} \textbf{fitted values} \qquad \widehat{y} &= B\widehat{\beta}_\lambda = B(B^{\mathrm{T}}B + \lambda S)^{-1}B^{\mathrm{T}}y = H_\lambda y, \\ \textbf{equivalent degrees of freedom} \quad \mathrm{df}_\lambda &= \mathrm{tr}(H_\lambda) = \sum_{k=1}^{K} \frac{1}{1 + \eta_k \lambda}, \end{aligned}$$

  where
  - $\eta_1 \le \cdots \le \eta_K \in [0, 1]$ are the eigenvalues of $(B^{\mathrm{T}}B)^{-1/2} S (B^{\mathrm{T}}B)^{-1/2}$,
  - $\eta_1 = \eta_2 = 0$, corresponding to the null space of S, so
  - df_λ is monotone decreasing in λ, with

$$(\lambda = 0) \qquad K \ge \mathrm{df}_\lambda \ge 2 \qquad (\lambda \to \infty).$$

Higher-order splines

- The p**th degree spline** basis with **knots** $x_1^* < \cdots < x_K^*$ is

$$1, x, \ldots, x^p, (x - x_1^*)_+^p, \ldots, (x - x_K^*)_+^p,$$

  where $u_+ = \max(u, 0)$ is the **positive part function**.
- The resulting basis 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** x_0^* and x_{K+1}^* and setting up an **augmented knot sequence**

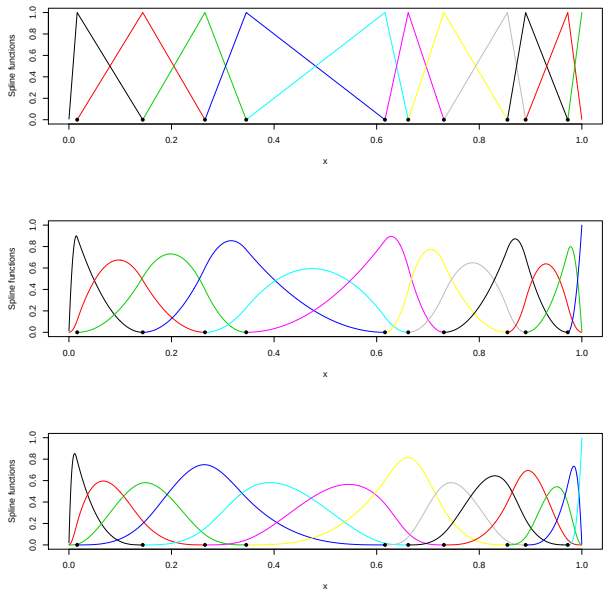
$$\tau_1 \leq \cdots \leq \tau_M \leq x_0^* \leq \tau_{M+1} = x_1^* \leq \cdots \leq \tau_{M+K} = x_K^* \leq x_{K+1}^* \leq \tau_{K+1+M} \leq \cdots \leq \tau_{K+2M};$$

  typically the τ_k outside $[x_0^*, x_{K+1}^*]$ are set to the boundary knot values. Then

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

  where we set $B_{k,1} \equiv 0$ if $\tau_k = \tau_{k+1}$ (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), and the τ_k set to equal the boundary knots.

Linear, quadratic and cubic B-splines



Natural cubic spline

☐ Suppose the x_j are distinct (no loss of generality) 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

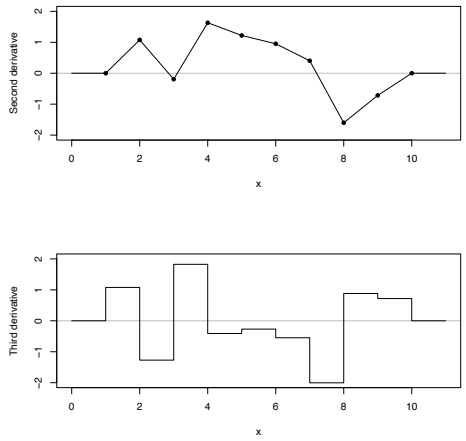
- has $K = n$ knots, at $x_1 < \cdots < x_n$,
- is a cubic polynomial on each interval between knots,
- is continuous, with continuous first and second derivatives at each knot, and
- is linear on $[a, x_1]$ and $[x_n, b]$, with zero second and third derivatives at x_1 and x_n,
- has

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

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 be constructed by integrating a linear second derivative function $\mu''(x)$ which is determined by $\mu''(x_2), \ldots, \mu''(x_{K-1})$ and because $\mu''(x) \equiv 0$ for $x \notin (x_1, x_K)$.

☐ On integrating twice we gain two constants: $\mu(x) = \beta_0 + \beta_1 x + \int_0^x \int_0^{x'} \mu''(u)\, \mathrm{d}u\, \mathrm{d}x'$.

☐ Above $x_1 = 1, \ldots, x_{10} = 10$, so the spline is determined by $\mu''(2), \ldots, \mu''(9)$ and the line.

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': i.e., 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 32 *Suppose $n \geq 2$, that $a < x_1 < \cdots < x_n < b$, and that μ is the natural cubic spline interpolating $y_1, \ldots, y_n$ 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 μ minimises the **roughness penalty** $\lambda \int_{\mathcal{X}} \mu''^2$ in a larger class of functions than that to which it belongs, making it a natural choice as an interpolant, because minimising

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

for $\tilde{\mu} \in \mathcal{S}_2(\mathcal{X})$ will automatically result in a natural cubic spline μ: if $\tilde{\mu}(x_j) = \mu(x_j)$, then the penalty is reduced by using μ.

Note to Theorem 32

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 $x \in \mathcal{X}$.

More splines

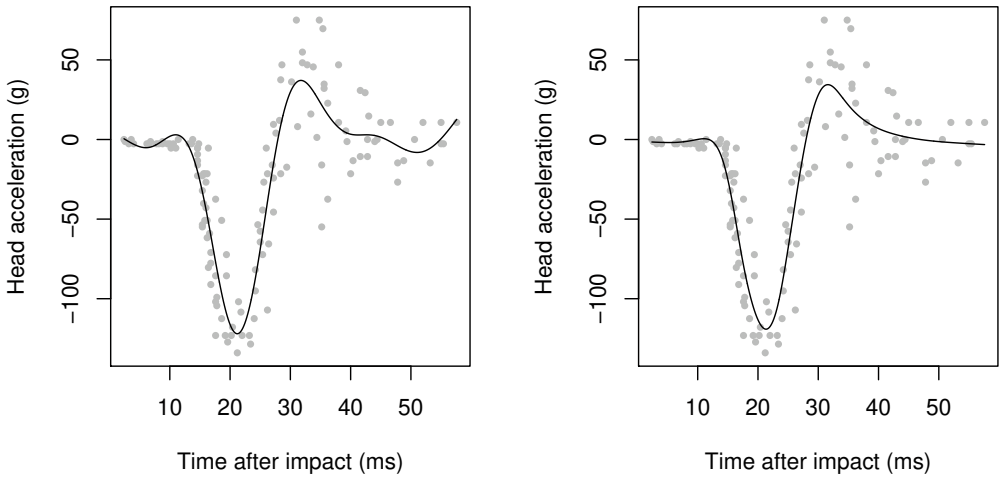
- Sometimes cyclic effects (e.g., seasonality, diurnal variation) must be modelled smoothly, so (e.g.) December joins smoothly onto January. Then the penalty and spline basis must be modified accordingly, to give a **cyclic (cubic) spline**.
- **P-splines** are a version of B-splines (usually with equally-spaced knots) in which a difference penalty is applied to the parameters to control the wiggliness of μ, e.g.,

$$\sum_{k=1}^{K-1} w_k(\beta_{k+1}-\beta_k)^2 = \beta^{\mathrm{T}} D^{\mathrm{T}} W D \beta, \quad \text{with} \quad D = \begin{pmatrix} -1 & 1 & 0 & 0\cdots & \\ 0 & -1 & 1 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{pmatrix},$$

  and $W = \mathrm{diag}(w_1, \ldots, w_{K-1})$. These are easy to set up and flexible, but messy if the knots are not equi-spaced, and the penalty is less readily interpreted.
- For an **adaptive spline** we can let $w_k \equiv w_k(x)$ vary with x, for example setting $w(x) = B(x)\lambda_{L\times 1}$ and thus having $D^{\mathrm{T}}WD = \sum_l \lambda_l D^{\mathrm{T}} \mathrm{diag}\{B_l(x)\} D$, where $B_l(x)$ is the lth column of $B(x)$, then estimating the vector λ.
- Other possibilities include (Wood, 2017, Chapter 5)
  - **shape-constrained splines** to impose, e.g., monotonicity on the fit;
  - **thin-plate**, **Duchon** and **tensor product** splines used in spatial problems; and
  - **soap film splines** used when smoothing over complex domains.

Motorcycle data: adaptive fit

Standard (left) and adaptive (right) spline fits, the latter with $K = 40$ and $L = 5$:



3.5 General Framework

slide 239

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 (later).
- To include ordinary covariates and allow for weights, we write

$$y \mid b \sim (B\theta, \sigma^2 W), \quad B\theta = X\beta + Zb,$$

where $B = (X, Z)$ is $n \times d$, $\theta = (\beta^{\mathrm{T}}, b^{\mathrm{T}})^{\mathrm{T}}$ is $d \times 1$, $d = p + q$ and

- the $n \times p$ matrix X represents the ordinary covariates, plus any unpenalised columns for smooth components,
- the $p \times 1$ parameter vector β is unpenalized,
- the $n \times q$ matrix Z represents the bases for any smooth functions,
- the $q \times 1$ vector b is penalized,
- the $n \times n$ diagonal matrix $W = \mathrm{diag}(w_1, \ldots, w_n)$ contains positive weights,

and everything 'goes through as before'.

Additivity and identifiability

- Consider the **additive model**

$$\mathrm{E}(y) = \mu_1(x) + \mu_2(z),$$

where μ_1, μ_2 belong to suitable classes of smooth functions; 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 $\mathcal{X}_2 \subset \mathbb{R}^2$.

- 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

$$\mathrm{E}(y) = Zb = \begin{pmatrix} Z_1(x) & Z_2(z) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

where we penalise the q_1 elements of b_1 and the q_2 elements of b_2.

Ensuring 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 $C_{a\times q} b_{q\times 1} = 0_{a\times 1}$, with $a < q$, write

$$C_{q\times a}^{\mathrm{T}} = Q_{q\times q} R_{q\times a} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix},$$

  where Q is orthogonal,
  - Q_1 has dimension $q \times a$,
  - Q_2 has dimension $q \times (q-a)$, and
  - R_1 has dimension $a \times a$ and is upper triangular.

  Then if we set $b_{q\times 1} = Q_2 b'_{(q-a)\times 1}$, 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 $(q-1)\times 1$.

Penalty formulation

- Minimise

$$(y - B\theta)^{\mathrm{T}} W (y - B\theta) + \theta^{\mathrm{T}} S_\lambda \theta = (y - X\beta - Zb)^{\mathrm{T}} W (y - X\beta - Zb) + \theta^{\mathrm{T}} S_\lambda \theta$$

  where S_λ is a sum of symmetric positive semi-definite $d \times d$ matrices S_m, such that

$$\theta^{\mathrm{T}} S_\lambda \theta = \theta^{\mathrm{T}} \left(\sum_{m=1}^{M} \lambda_m S_m \right) \theta = \sum_{m=1}^{M} \lambda_m b_m^{\mathrm{T}} S_m^* b_m, \quad \lambda_m \geq 0,$$

  where S_m^* is the non-zero diagonal block of S_m and b has sub-vectors $b_1, \ldots, b_M$.
- With $M = 2$, β, b_1 and b_2 are vectors of respective lengths p, q_1 and q_2, and S_1^* and S_2^* are square matrices of sides q_1 and q_2, so

$$\theta = \begin{pmatrix} \beta \\ b_1 \\ b_2 \end{pmatrix}, \quad S_\lambda = \lambda_1 S_1 + \lambda_2 S_2 = \lambda_1 \begin{pmatrix} 0 & 0 & 0 \\ 0 & S_1^* & 0 \\ 0 & 0 & 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S_2^* \end{pmatrix},$$

  with S_1 and S_2 partitioned conformably with θ.
- Let S_λ^* denote the $q \times q$ corner of S_λ corresponding to b; here $S_\lambda^* = \mathrm{diag}(\lambda_1 S_1^*, \lambda_2 S_2^*)$.
- Note that $|S_\lambda|_+ = |S_\lambda^*|_+$.

Estimation

- For fixed λ, the minimiser and fitted values for

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

are

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W y, \quad \widehat{y}_\lambda = B\widehat{\theta}_\lambda = B(B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W y = H_\lambda y.$$

- If the unpenalized least squares estimator $\widehat{\theta} = (B^{\mathrm{T}} W B)^{-1} B^{\mathrm{T}} W y$ exists, then

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}} W B + S_\lambda)^{-1} B^{\mathrm{T}} W B \widehat{\theta} = \widehat{\theta} - (B^{\mathrm{T}} W B + S_\lambda)^{-1} S_\lambda \widehat{\theta} = P_\lambda \widehat{\theta},$$

and if $\widehat{y}$ is the unpenalised fitted value, then

$$\widehat{y}_\lambda = \widehat{y} - B(B^{\mathrm{T}} W B + S_\lambda)^{-1} S_\lambda \widehat{\theta}.$$

- Now we must decide
  - how many degrees of freedom for each smooth?
  - how to select the smoothing parameters?

Amount of smoothing

- We write

$$\widehat{\theta}_\lambda = P_\lambda \widehat{\theta},$$

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 193 with just one smooth term we defined

$$\mathrm{edf}_\lambda = \mathrm{tr}(H_\lambda) = \mathrm{tr}(P_\lambda) = \sum_{r=1}^{p+q} P_{\lambda,rr} \in (p, p+q),$$

which gives the usual definition for a linear model.

- If $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, b_1^{\mathrm{T}}, \ldots, b_M^{\mathrm{T}})$, we define the **effective degrees of freedom** edf_{λ_m} associated to the mth smooth as being the sum of those $P_{\lambda,rr}$ that correspond to the elements of b_m in θ.
- To choose the vector λ we use either
  - $\mathrm{CV}(\lambda)$ or $\mathrm{GCV}(\lambda)$ (second-order assumptions),
  - REML (normal-theory assumptions).
- Must optimise over (log) λ, e.g., by grid search (CV/GCV) or other methods (REML).

Inference

- So far we have discussed only 'point estimation' of a smooth function $\mu(x)$, but in applications we also want
  - pointwise confidence intervals for smooth functions,
  - overall confidence bands for (say) $\{\mu(x) : x \in \mathcal{S}\}$, 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?'
- Under the normal model we have the Bayesian interpretation from slide 191,

$$\theta \mid y, \sigma^2, \lambda \sim \mathcal{N}_d\left(\widehat{\theta}_\lambda, V_\lambda\right), \quad V_\lambda = \sigma^2(B^{\mathrm{T}}WB + S_\lambda)^{-1},$$

  from which we can simulate to find bounds for any function $A(\theta)$.
- If $A(\theta) = A_{m\times d}\theta$, then

$$A\theta \mid y, \sigma^2, \lambda \sim \mathcal{N}_m(A\widehat{\theta}_\lambda, AV_\lambda A^{\mathrm{T}}),$$

  and generalisation of (10) gives that its mean square error is

$$\mathrm{MSE} = \mathrm{E}\left(\|A\widehat{\theta}_\lambda - A\theta\|^2\right) = \mathrm{tr}(AV_\lambda A^{\mathrm{T}}),$$

  which takes into account both estimation error and prior uncertainty about θ.

Average coverage probabilities

- Bayesian credible intervals have good frequentist properties, averaged over the domain of x.
- Let the random index variable J choose the m rows a_j^{T} of A with equal probabilities, and aim to choose constants d and c_j such that the **average coverage probability**

$$\mathrm{ACP} = \mathrm{P}\left\{|a_J^{\mathrm{T}}\widehat{\theta}_\lambda - a_J^{\mathrm{T}}\theta| \le dc_J\right\} = 1 - \alpha;$$

  i.e., ACP has a desired value averaged over y, θ and J.
- The random variable

$$a_J^{\mathrm{T}}(\widehat{\theta}_\lambda - \theta)/c_J = a_J^{\mathrm{T}}\{\widehat{\theta}_\lambda - \mathrm{E}(\widehat{\theta}_\lambda)\}/c_J + a_J^{\mathrm{T}}\{\mathrm{E}(\widehat{\theta}_\lambda) - \theta\}/c_J = S + T,$$

  say, has a mixture of normal distributions, where
  - S is approximately normal and $\mathrm{E}(S) = 0$,
  - T is random (because of J) with $\mathrm{E}(T) \approx 0$, but $\mathrm{var}(T) \ll \mathrm{var}(S)$.
- We now choose $C = \mathrm{diag}(c_1, \ldots, c_m) = \mathrm{diag}(AV_\lambda A^{\mathrm{T}})^{1/2}$, so that

$$\mathrm{var}(S + T) \approx m^{-1}\mathrm{E}\left\{\|C^{-1}A(\widehat{\theta}_\lambda - \theta)\|^2\right\} = m^{-1}\mathrm{tr}\left(C^{-1}AV_\lambda A^{\mathrm{T}}C^{-1}\right) = 1,$$

  and then setting $d = z_{1-\alpha/2}$ gives the required value for ACP.
- This ignores estimation error for σ^2 and λ.

Example: Average coverage probability

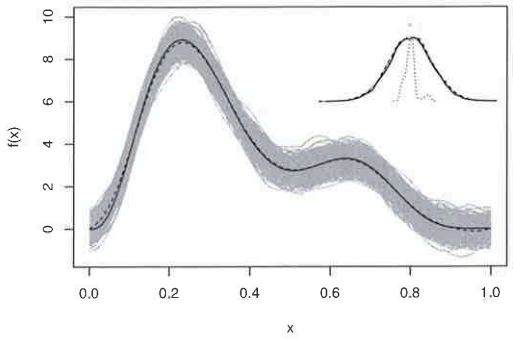
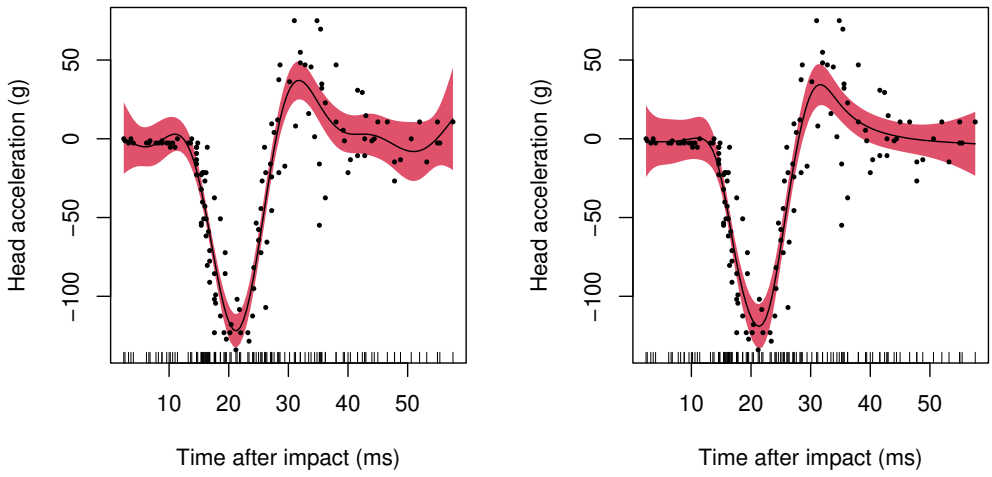


Figure 6.7 *The Nychka (1988) idea. The main black curve shows a true function $f(x)$, while the grey curves show 500 replicate spline estimates $\hat{f}(x)$. The dashed curve is $\mathbb{E}\hat{f}(x)$. Inset at top* ***right are*** *scaled* ***kernel smooth*** *estimates of the distributions of the sampling error, $\hat{f} - \mathbb{E}\hat{f}$ (****continuous*** *black); the* ***bias,*** *$\mathbb{E}\hat{f} - f$, evaluated at a random x (dotted) and $\hat{f} - f$ evaluated at a random x (dashed). In grey is the normal approximation to the dashed curve. Evaluation at a random x turns the bias into a random variable, which has substantially lower variance than the approximately normal $\hat{f} - \mathbb{E}\hat{f}$. Hence the sum of the randomized bias and sampling error is approximately normally distributed. The variance of this sum turns out to be well approximated by the Bayesian posterior covariance for f.*

(Wood, 2017)

Example: Motorcycle data

Standard (left) and adaptive (right) spline fits, the latter with $K = 40$ and $L = 5$, and 95% pointwise confidence intervals:

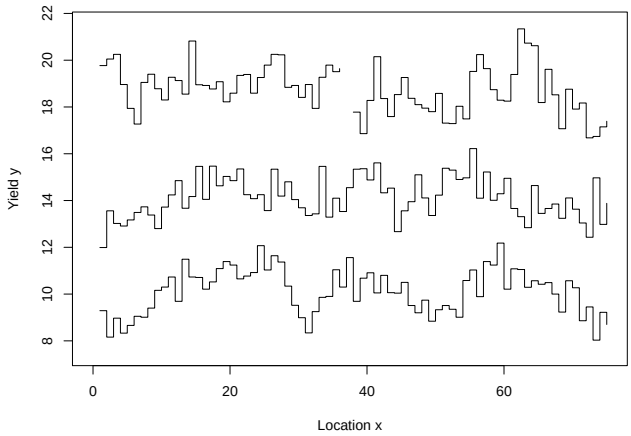


Example: Spring barley data

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

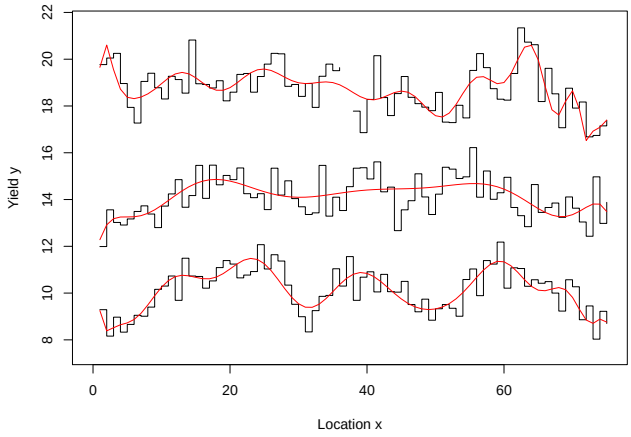
Location x	Block 1		Block 2		Block 3	
	Variety	Yield y	Variety	Yield y	Variety	Yield y
1	57	9.29	49	7.99	63	11.77
2	39	8.16	18	9.56	38	12.05
3	3	8.97	8	9.02	14	12.25
4	48	8.33	69	8.91	71	10.96
5	75	8.66	29	9.17	22	9.94
6	21	9.05	59	9.49	46	9.27
7	66	9.01	19	9.73	6	11.05
8	12	9.40	39	9.38	30	11.40
9	30	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.

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.

Example: Spring barley data

- We fit a model with parametric variety effects and smooth effects for the fertility patterns in the blocks,

$$y_{n\times 1} \sim (X_{n\times 75}\beta_{75\times 1} + Z_1 b_1 + Z_2 b_2 + Z_3 b_3, \sigma^2 I_n),$$

  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
  - b_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 255).

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)
```

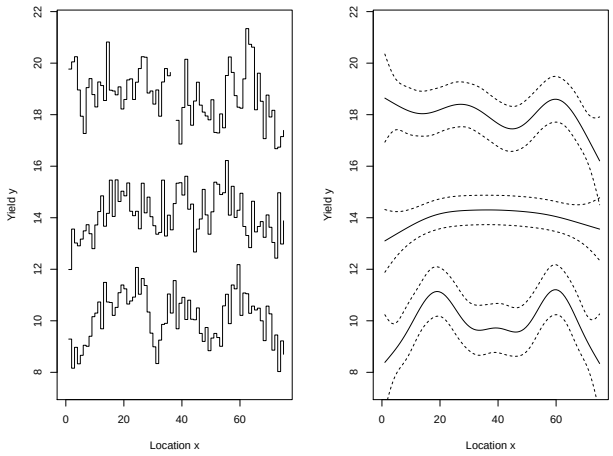

Example: Spring barley data

- 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

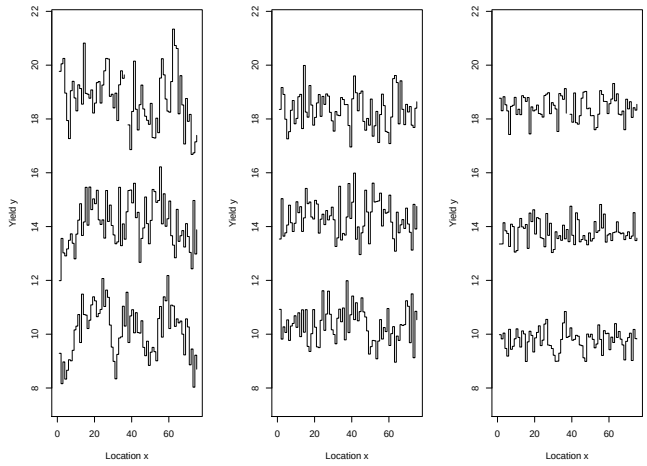
- The $P_{\lambda,rr}$ need not be positive, though their total for each smooth is positive.
- In applications it would be wise to check whether increasing q_m would lead to very different fits.

Example: Spring barley data



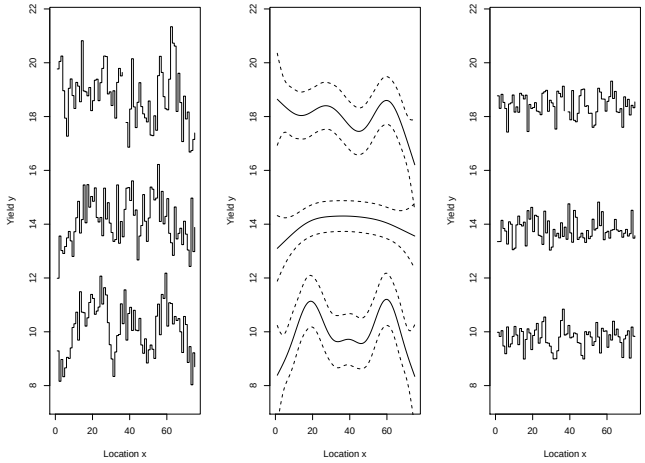
Left: data (offset by adding 4 and 8 to blocks 2 and 3).
Right: estimated fertility patterns (with estimated df 7.2, 3, 6.1) and 95% unconditional pointwise confidence intervals, fitted using REML. The intervals are wider for blocks 1 and 3.

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 $\lambda_\beta S_\beta$ and minimise the penalised sum of squares

$$(y - B\theta)^{\mathrm{T}}(y - B\theta) + \theta^{\mathrm{T}} S_\lambda \theta,$$

where

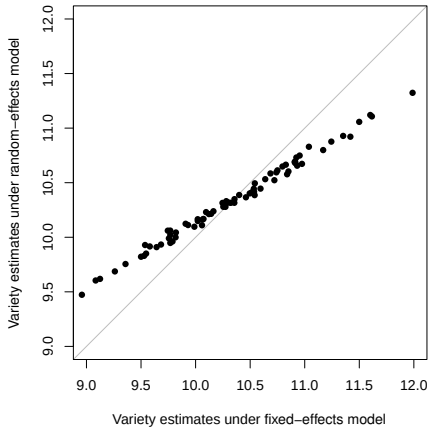
$$S_\lambda = \lambda_\beta S_\beta + \lambda_1 S_1 + \lambda_2 S_2 + \lambda_3 S_3,$$

where $S_\beta = \mathrm{diag}(I_{75}, 0)$.

- ☐ The effective degrees of freedom are then 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:



Comments

- Penalised estimation extends the basic smoothers to include
  - parametric terms in models,
  - several smooth terms,
  - spatial and more complex smoothing,
  - 'random effect' parameters,

  and extends to **generalized additive models** in a natural way.
- The baseline variance σ^2 and smoothing parameter(s) λ are estimated using cross-validation under second-order assumptions or REML under normality.
- The empirical Bayes formulation allows inference on parameters and smooth functions in a unified way — usually ignoring the uncertainty for σ^2 and λ is not too critical.
- In practice n and d may be very big, so direct matrix inversion is computationally painful, and then indirect methods (e.g., based on the Woodbury formula) are needed to compute $\widehat{\theta}_\lambda$ and V_λ.

3.6 Components of Variance

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 population mean diastolic blood pressure (DBP),
  - b_p is the difference between the patient and population mean DBP,
  - e_{pv} is the difference between this and the 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,

  and

$$b_p \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2) \perp\!\!\!\perp e_{pv} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2) \perp\!\!\!\perp \varepsilon_{pvm} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

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

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

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

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

- 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
```


Nested model ANOVA

- ☐ For the nested model

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

and with a dot and bar denoting averaging over that index, we write

$$y_{pvm} - \overline{y}_{...} = y_{pvm} - \overline{y}_{pv\cdot} + \overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot} + \overline{y}_{p\cdot\cdot} - \overline{y}_{...},$$

and note that

$$\begin{aligned} y_{pvm} - \overline{y}_{pv\cdot} &= \varepsilon_{pvm} - \overline{\varepsilon}_{pv\cdot}, \\ \overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot} &= e_{pv} + \overline{\varepsilon}_{pv\cdot} - (\overline{e}_{p\cdot} + \overline{\varepsilon}_{p\cdot\cdot}), \\ \overline{y}_{p\cdot\cdot} - \overline{y}_{...} &= b_p + \overline{e}_{p\cdot} + \overline{\varepsilon}_{p\cdot\cdot} - (\overline{b}_{\cdot} + \overline{e}_{\cdot\cdot} + \overline{\varepsilon}_{...}), \end{aligned}$$

so the overall sum of squares is

$$\begin{aligned} \sum_{p,v,m} (y_{pvm} - \overline{y}_{...})^2 &= \sum_{p,v,m} (y_{pvm} - \overline{y}_{pv\cdot})^2 + \sum_{p,v,m} (\overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot})^2 + \sum_{p,v,m} (\overline{y}_{p\cdot\cdot} - \overline{y}_{...})^2 \\ &= \sum_{p,v,m} (y_{pvm} - \overline{y}_{pv\cdot})^2 + M \sum_{p,v} (\overline{y}_{pv\cdot} - \overline{y}_{p\cdot\cdot})^2 + VM \sum_{p} (\overline{y}_{p\cdot\cdot} - \overline{y}_{...})^2, \end{aligned}$$

where these terms are independent sums of squares for variables that are

$$\mathcal{N}(0, \sigma^2), \quad \mathcal{N}(0, \sigma_e^2 + \sigma^2/M), \quad \mathcal{N}\{0, \sigma_b^2 + \sigma_e^2/V + \sigma^2/(VM)\}.$$

Nested model ANOVA II

- Hence

$$\begin{aligned}\sum_{p,v,m}(y_{pvm}-\overline{y}_{pv\cdot})^2 &\sim \sigma^2\chi^2_{PV(M-1)},\\ \sum_{p,v,m}(\overline{y}_{pv\cdot}-\overline{y}_{p\cdot\cdot})^2 &\sim M(\sigma_e^2+\sigma^2/M)\chi^2_{P(V-1)} \stackrel{\mathrm{D}}{=} (M\sigma_e^2+\sigma^2)\chi^2_{P(V-1)},\\ \sum_{p,v,m}(\overline{y}_{p\cdot\cdot}-\overline{y}_{\cdot\cdot\cdot})^2 &\sim VM\left(\sigma_b^2+\tfrac{\sigma_e^2}{V}+\tfrac{\sigma^2}{VM}\right)\chi^2_{P-1} \stackrel{\mathrm{D}}{=} (VM\sigma_b^2+M\sigma_e^2+\sigma^2)\chi^2_{P-1},\end{aligned}$$

  and we can estimate the components of variance σ^2, σ_e^2 and σ_b^2 from the ANOVA table.
- The interpretation of the ANOVA depends on whether we regard $\delta_b^2=\sum_p(b_p-\overline{b}_\cdot)^2$ and $\delta_e^2=\sum_{p,v}(e_{pv}-\overline{e}_{p\cdot})^2$ as random or fixed:

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

Nested and crossed ANOVA

- 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.

- Different comparisons require appropriate baseline variances:
  - if we are interested in how patient p's response varies from visit to visit, we use

$$\overline{y}_{pv_1\cdot} - \overline{y}_{pv_2\cdot} = \mu + b_p + e_{pv_1} + \overline{\varepsilon}_{pv_1\cdot} - (\mu + b_p + e_{pv_2} + \overline{\varepsilon}_{pv_2\cdot}) \sim \mathcal{N}(0, 2\sigma_e^2 + 2\sigma^2/M),$$

as a basis for a test of a significant difference, whereas to compare average blood pressures for two different patients we use

$$\overline{y}_{p_1\cdot\cdot} - \overline{y}_{p_2\cdot\cdot} = b_{p_1} + \overline{e}_{p_1\cdot} + \overline{\varepsilon}_{p_1\cdot\cdot} - (b_{p_2} + \overline{e}_{p_2\cdot} + \overline{\varepsilon}_{p_2\cdot\cdot}) \sim \mathcal{N}\{0, 2\sigma_b^2 + 2\sigma_e^2/V + 2\sigma^2/(VM)\}.$$

- **Split-unit designs** are set up to make the most important comparisons within units (here patients) and less important ones between units, and the ANOVA reflects this.

General form

- We could have written the nested model above as

$$y = 1_n\mu + X_b b + X_e e + \varepsilon,$$

with design matrices X_b and X_e for the patient and visit-within-patient effects.

- Then if
  - b and e are treated as fixed (ordinary parameters),

$$y \sim \mathcal{N}_n(1_n\mu + X_b b + X_e e, \sigma^2 I_n),$$

  - b is treated as fixed but $e \sim \mathcal{N}_{PV}(0, \sigma_e^2 I_{PV})$, then

$$y \sim \mathcal{N}_n(1_n\mu + X_b b, \sigma_e^2 X_e X_e^{\mathrm{T}} + \sigma^2 I_n),$$

  - and if $b \sim \mathcal{N}_P(0, \sigma_b^2 I_P)$ independent of $e \sim \mathcal{N}_{PV}(0, \sigma_e^2 I_{PV})$, then

$$y \sim \mathcal{N}_n(1_n\mu, \sigma_b^2 X_b X_b^{\mathrm{T}} + \sigma_e^2 X_e X_e^{\mathrm{T}} + \sigma^2 I_n).$$

- Hence random e or b give patterned covariance matrices depending on their variances.

Summary

- ☐ Components of variance ANOVA is easily performed directly for 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 — hence **mixed models**, i.e., models with both random and fixed parts, arising in many different settings (and with different names):
  - components of variance (as above),
  - 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, ...

3.7 Linear Mixed Model

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 $\Omega = \sigma^2 I_n$.
- ☐ This has the same structure as when smoothing, with the columns of Z giving the structure of the random effects.
- ☐ 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 the vector of distinct variance ratios appearing in Δ^{-1} (e.g., $\sigma_b^2/\sigma^2, \ldots$).
- ☐ 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 avoid working with Δ.

Maximum likelihood estimation

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

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

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 maximum likelihood estimation

☐ 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 $\mathcal{N}_q(\tilde{b}, H^{-1})$ density.

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
$$\begin{pmatrix} \mathrm{var}(\widehat{\beta})_{p\times p} & \cdot \\ \cdot & \cdot \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}\Omega^{-1}X & X^{\mathrm{T}}\Omega^{-1}Z \\ Z^{\mathrm{T}}\Omega^{-1}X & Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1} \end{pmatrix}^{-1}_{d\times d},$$
where $d = p + q$, which involves only $O\{nd^2\}$ 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 are 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):
$$\begin{aligned} \mathrm{E}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}\left(y - X\beta\right), \\ \mathrm{var}(b \mid y) &= \left(Z^{\mathrm{T}}\Omega^{-1}Z + \Omega_b^{-1}\right)^{-1}. \end{aligned}$$
- 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

$$\begin{aligned} \mathrm{E}(b \mid y) &= \Omega_b Z^{\mathrm{T}} \left\{\Omega^{-1} - \Omega^{-1}Z\left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1}Z\right)^{-1} Z^{\mathrm{T}}\Omega^{-1}\right\} (y - X\beta) \\ &= \Omega_b \left\{I_q - Z^{\mathrm{T}}\Omega^{-1}Z\left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1}Z\right)^{-1}\right\} Z^{\mathrm{T}}\Omega^{-1}(y - X\beta) \\ &= \Omega_b \left\{\Omega_b^{-1}\left(\Omega_b^{-1} + Z^{\mathrm{T}}\Omega^{-1}Z\right)^{-1}\right\} Z^{\mathrm{T}}\Omega^{-1}(y - X\beta), \end{aligned}$$

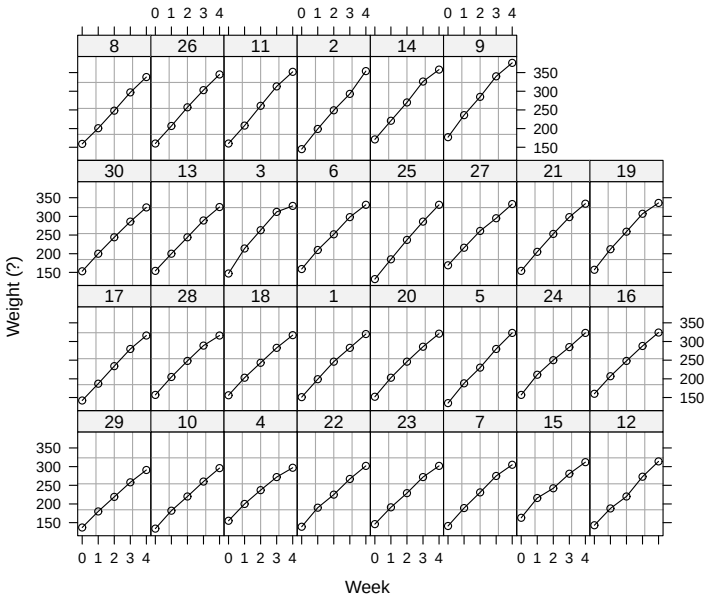
as required, where we wrote the term in braces in the second line as $I - B(A + B)^{-1} = A(A + B)^{-1}$, with $A = \Omega_b^{-1}$ and $B = Z^{\mathrm{T}}\Omega^{-1}Z$.

Example: 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

Example 33 (Rat growth data)

☐ *Write*

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

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,*

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

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

☐ *We set* $(x_{j1}, \ldots, x_{j5}) = (0, \ldots, 4)$*, 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

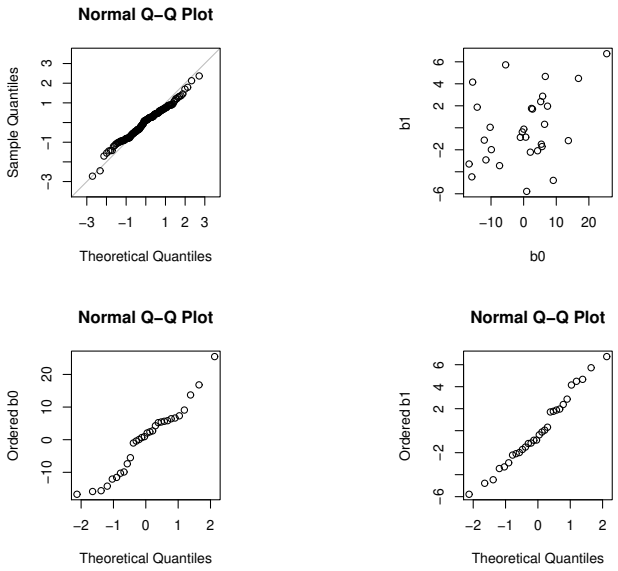
Results from fit of mixed model to rat growth data, using REML. Values in parentheses are for ML fit. In each case $\widehat{\sigma}^2 = 5.82^2$.

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 involves tests on the boundary of the parameter space, which have nasty asymptotic properties: if $\psi = 0$, then a likelihood ratio statistic for testing $\psi = 0$ satisfies $W \dot{\sim} \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ as $n \to \infty$, meaning that

$$\mathrm{P}_0(W = 0) = \tfrac{1}{2}, \quad \mathrm{P}_0(W > w) = \tfrac{1}{2}\mathrm{P}(\chi^2_1 > w), \quad w > 0.$$

  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 nastier computations.

3.8 Generalized Additive Models

Generalized additive model

- Now we write

$$\mathrm{E}(y) = \mu, \quad g(\mu) = \eta = B\theta = X\beta + Zb,$$

  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,

$$\ell_\lambda(\theta) = \ell(\theta) - \tfrac{1}{2}\theta^{\mathrm{T}} S_\lambda \theta = \sum_{j=1}^{n} \ell_j\{\eta_j(\theta)\} - \tfrac{1}{2}\theta^{\mathrm{T}} S_\lambda \theta,$$

  where $\theta_{d\times 1}$ (with $d = p + q$) contains $\beta_{p\times 1}$ and $b_{q\times 1}$, the latter penalized using a symmetric positive semidefinite $d \times d$ matrix S_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

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}Wz,$$

where S_λ is the penalty matrix, and

$$\begin{aligned} B_{n\times d} &= \partial\eta/\partial\theta^{\mathrm{T}}, \quad \text{(design matrix)} \\ W_{n\times n} &= \mathrm{diag}(w_1,\ldots,w_n), \quad w_j = \{\mathrm{E}(-\partial^2\ell_j/\partial\eta_j^2)\}, \quad \text{(weights)} \\ u_{n\times 1} &= \partial\ell/\partial\eta, \quad \text{(score vector)}, \\ z_{n\times 1} &= B\theta + W^{-1}u, \quad \text{(adjusted dependent variable)}. \end{aligned}$$

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 $\widehat{\theta}_\lambda$ we use the **PIWLS algorithm**:

☐ fix λ and take an initial $\widehat{\theta}_\lambda$. Repeat

- compute η, B, W, u, z;
- compute new $\widehat{\theta}_\lambda = (B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}Wz$;

until changes in $\ell_\lambda(\widehat{\theta}_\lambda)$ (or $\widehat{\theta}_\lambda$, or both) are lower than some tolerance.

☐ We may add a line search: if $\ell_\lambda(\widehat{\theta}_{\lambda,\mathrm{new}}) < \ell_\lambda(\widehat{\theta}_{\lambda,\mathrm{old}})$, halve the step length and try again.

Note: Derivation of PIWLS algorithm

☐ To find the estimate $\widehat{\theta}_\lambda$ starting from a trial value θ, we make a Taylor series expansion in the score equation

$$0 = \frac{\partial\ell_\lambda(\widehat{\theta}_\lambda)}{\partial\theta} \doteq \frac{\partial\ell_\lambda(\theta)}{\partial\theta} + \frac{\partial^2\ell_\lambda(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}(\widehat{\theta}_\lambda - \theta),$$

where

$$\frac{\partial\ell_\lambda(\theta)}{\partial\theta} = B^{\mathrm{T}}u(\theta) - S_\lambda\theta, \quad \frac{\partial^2\ell_\lambda(\theta)}{\partial\theta_r\partial\theta_s} = \sum_{j=1}^{n}\frac{\partial\eta_j(\theta)}{\partial\theta_r}\frac{\partial^2\ell_j(\theta)}{\partial\eta_j^2}\frac{\partial\eta_j(\theta)}{\partial\theta_s} + \sum_{j=1}^{n}\frac{\partial^2\eta_j(\theta)}{\partial\theta_r\partial\theta_s}u_j(\theta) + S_{\lambda,r,s},$$

where $B \equiv B(\theta) = \partial\eta/\partial\theta^{\mathrm{T}}$. If we use the approximation

$$-\frac{\partial^2\ell_\lambda(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} \doteq B^{\mathrm{T}}WB + S_\lambda, \quad W = \mathrm{diag}\left\{-\mathrm{E}\left(\partial^2\ell_j/\partial\eta_j^2\right)\right\},$$

where the diagonal matrix of second derivatives is replaced by its expectation, then

$$\begin{aligned} 0 &\doteq B^{\mathrm{T}}u(\theta) - S_\lambda\theta - (B^{\mathrm{T}}WB + S_\lambda)(\widehat{\theta}_\lambda - \theta) \\ &= B^{\mathrm{T}}u(\theta) + B^{\mathrm{T}}WB\theta - (B^{\mathrm{T}}WB + S_\lambda)\widehat{\theta}_\lambda. \end{aligned}$$

If $B^{\mathrm{T}}WB + S_\lambda$ is invertible, this gives

$$\widehat{\theta}_\lambda \doteq (B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}(u + WB\theta) = (B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}Wz,$$

where $z = B\theta + W^{-1}u$, as required.

Relation with least squares

- With fixed λ, the penalized MLE

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}Wz$$

results from fixing θ, and then iteratively solving the minimization problem

$$\min_\theta \left\| \begin{pmatrix} W^{1/2}z \\ 0 \end{pmatrix}_{(n+d)\times 1} - \begin{pmatrix} W^{1/2}B \\ Q_\lambda \end{pmatrix}_{(n+d)\times d} \theta_{d\times 1} \right\|^2,$$

where Q_λ is a matrix square root of S_λ, i.e., $Q_\lambda^{\mathrm{T}}Q_\lambda = S_\lambda$.

- The corresponding smoothing matrix is taken to be

$$H_\lambda = B(B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}W,$$

and the effective degrees of freedom for a smooth component are defined as the sum of the corresponding diagonal elements of

$$P_\lambda = (B^{\mathrm{T}}WB + S_\lambda)^{-1}B^{\mathrm{T}}WB,$$

with both H_λ and P_λ evaluated at the final step of the iteration.

Approaches to iteration

- Having chosen how to choose λ for fixed θ, there are two main algorithms:
  - **performance iteration** — repeat { fix λ, update θ with one step of PIWLS, update λ } to convergence;
  - **outer iteration** — repeat { fix λ, iterate PIWLS to convergence, update λ } to convergence.
- Performance iteration
  - can be faster,
  - but since the objective function for θ 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 34 *Let $h(u)$ be a smooth convex function defined for $u \in \mathbb{R}^d$, 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*

$$I_n = \int_{\mathbb{R}^d} e^{-nh(u)}\,\mathrm{d}u.$$

Then $I_n = \tilde{I}_n\left\{1 + O(n^{-1})\right\}$, and its **Laplace approximation** *is*

$$\tilde{I}_n = \frac{(2\pi)^{d/2}}{|nh_2|^{1/2}} e^{-nh(\tilde{u})}.$$

- For marginal density approximation we let $\theta = (\beta_{p\times 1}^{\mathrm{T}}, b_{q\times 1}^{\mathrm{T}})^{\mathrm{T}} \sim \mathcal{N}_d(0, S_\lambda^-)$, and write

$$f(y;\beta,\lambda) = \int f(y;\theta) f(\theta;\lambda)\,\mathrm{d}\theta = \frac{|S_\lambda|_+^{1/2}}{(2\pi)^{d/2}} \int \exp\left\{\ell_\lambda(\theta)\right\}\,\mathrm{d}\theta,$$

  where β is unpenalised, $|S_\lambda|_+$ is the product of the non-negative eigenvalues of S_λ, and

$$\ell_\lambda(\theta) = \ell(\theta) - \tfrac{1}{2}\theta^{\mathrm{T}} S_\lambda \theta = O(n);$$

  the assumptions of Lemma 34 should be satisfied by $h(u) \equiv -n^{-1}\ell_\lambda(\theta)$.

Note on Lemma 34

☐ Close to $\tilde{u}$ a Taylor series expansion gives

$$h(u) \doteq h(\tilde{u}) + h'(\tilde{u})^{\mathrm{T}}(u-\tilde{u}) + \tfrac{1}{2}(u-\tilde{u})^{\mathrm{T}}h''(\tilde{u})(u-\tilde{u}) = h(\tilde{u}) + \tfrac{1}{2}(u-\tilde{u})^{\mathrm{T}}h_2(u-\tilde{u})$$

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),

$$\begin{aligned} I_n &\doteq e^{-nh(\tilde{u})} \int e^{-n(u-\tilde{u})^{\mathrm{T}}h_2(u-\tilde{u})/2}\,\mathrm{d}u \\ &= e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-z^2/2}\frac{\mathrm{d}u}{\mathrm{d}z}\,\mathrm{d}z \\ &= \left(\frac{(2\pi^d}{|nh_2|}\right)^{1/2} e^{-nh(\tilde{u})}, \end{aligned}$$

because the d-dimensional normal density has unit integral.

☐ A more detailed accounting is needed to get the error term. Take the scalar case ($d = 1$) for simplicity. 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. The same argument works for $m > 1$, but it is more of a bloodbath.

Comments on Laplace approximations

- ☐ 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 hessian matrix h_2 at $\tilde{u}$, so is easily found, numerically if necessary;
- ☐ the series is asymptotic, so the partial sums may not converge, and including additional terms 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 are far enough away from $\tilde{u}$;
- ☐ 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.

Approximate REML

- ☐ Laplace approximation gives the approximate restricted log likelihood

$$\ell_{\mathrm{p}}(\lambda) \equiv \tfrac{1}{2}\log|S_\lambda|_+ - \tfrac{1}{2}\log|B^{\mathrm{T}}WB^{\mathrm{T}} + S_\lambda| + \ell(\widehat{\theta}_\lambda) - \tfrac{1}{2}\widehat{\theta}_\lambda^{\mathrm{T}} S_\lambda \widehat{\theta}_\lambda + O_p(n^{-1}),$$

where $O_p(n^{-1})$ is a (random) term of order n^{-1} and

$$\widehat{\theta}_\lambda = (B^{\mathrm{T}}WB + S_\lambda)^{-1} B^{\mathrm{T}} W z$$

results from iterating PIWLS to convergence for fixed λ and satisfies $\partial \ell_\lambda(\widehat{\theta}_\lambda)/\partial\theta = 0$.

- ☐ The expression for $\widehat{\theta}_\lambda$ contains

$$B \equiv B(\widehat{\theta}_\lambda), \quad W \equiv W(\widehat{\theta}_\lambda), \quad z = B(\widehat{\theta}_\lambda)\widehat{\theta}_\lambda + W^{-1}(\widehat{\theta}_\lambda)u(\widehat{\theta}_\lambda),$$

which involve the first two derivatives of the log likelihood contributions ℓ_j.

- ☐ Newton–Raphson maximization of $\ell_{\mathrm{p}}(\lambda)$ requires its first two derivatives, so we need

$$\frac{\partial \widehat{\theta}_\lambda}{\partial \lambda}, \quad \frac{\partial^2 \widehat{\theta}_\lambda}{\partial \lambda \partial \lambda^{\mathrm{T}}},$$

which will involve the third and fourth derivatives of the ℓ_j . . . could be painful.

- ☐ A version of this is implemented in `mgcv`.

UK monthly AIDS reports 1983–1992

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end
Year	Quarter	$0^\dagger$	1	2	3	4	5	6	$\cdots$	$\geq$14	of 1992
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1988	1	31	80	16	9	3	2	8	$\cdots$	6	174
	2	26	99	27	9	8	11	3	$\cdots$	3	211
	3	31	95	35	13	18	4	6	$\cdots$	3	224
	4	36	77	20	26	11	3	8	$\cdots$	2	205
1989	1	32	92	32	10	12	19	12	$\cdots$	2	224
	2	15	92	14	27	22	21	12	$\cdots$	1	219
	3	34	104	29	31	18	8	6	$\cdots$		253
	4	38	101	34	18	9	15	6	$\cdots$		233
1990	1	31	124	47	24	11	15	8	$\cdots$		281
	2	32	132	36	10	9	7	6	$\cdots$		245
	3	49	107	51	17	15	8	9	$\cdots$		260
	4	44	153	41	16	11	6	5	$\cdots$		285
1991	1	41	137	29	33	7	11	6	$\cdots$		271
	2	56	124	39	14	12	7	10	$\cdots$		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

- Chain-ladder model: number of reports in row j and column k is Poisson, with mean

$$\mu_{jk} = \exp(\alpha_j + \beta_k),$$

  but
  - why should there be different parameters α_j and β_k for every row and column?
  - Wouldn't smooth variation be more plausible?
- Better models (maybe?):

$$\mu_{jk} = \exp\{s(j) + \beta_k\}, \quad \mu_{jk} = \exp\{s(j) + s(k)\},$$

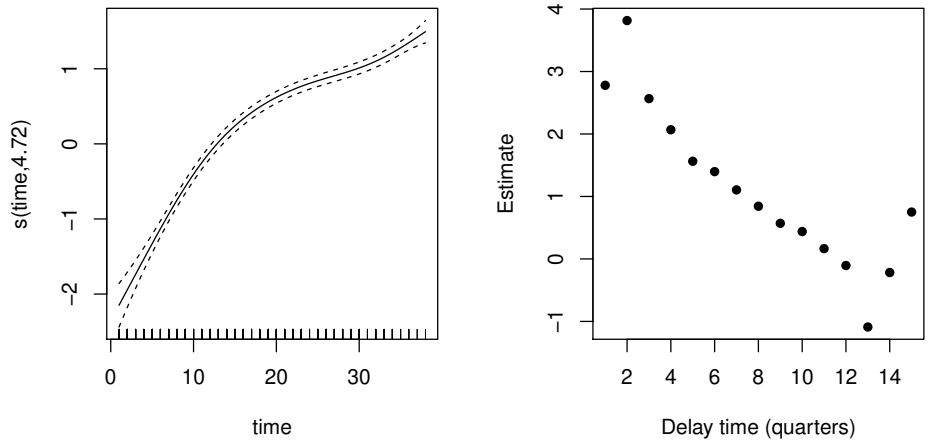
  where the time effect $s(j)$ and the delay effect $s(k)$ vary smoothly.
- Should also account for the overdispersion ...

Example: AIDS data

```
library(mgcv); library(boot)
data(aids)
aids.in <- aids[c(1:570)[as.logical(1-aids$dud)],] # these are elements in the two-way table
aids.glm <- glm(y~factor(time)+factor(delay),family=quasipoisson,data=aids.in)
aids.gam1 <- mgcv::gam(y~s(time,k=20)+factor(delay)-1,family=quasipoisson,data=aids.in)
plot(aids.gam1,page=1)
> anova(aids.gam1)

Formula:
y ~ s(time, k = 20) + factor(delay)

Parametric Terms:
              df     F p-value
factor(delay) 14 261.6  <2e-16

Approximate significance of smooth terms:   # Ref.df can be ignored
          edf Ref.df     F p-value
s(time) 4.891  6.129 189.1  <2e-16

aids.gam2 <- mgcv::gam(y~s(time,k=20)+s(delay,k=15),family=quasipoisson,data=aids.in)
> anova(aids.gam2)

Formula:
y ~ s(time, k = 20) + s(delay, k = 15)

Approximate significance of smooth terms:
            edf Ref.df     F p-value
s(time)   4.896  6.134 189.0  <2e-16
s(delay) 11.453 12.754 285.5  <2e-16
```

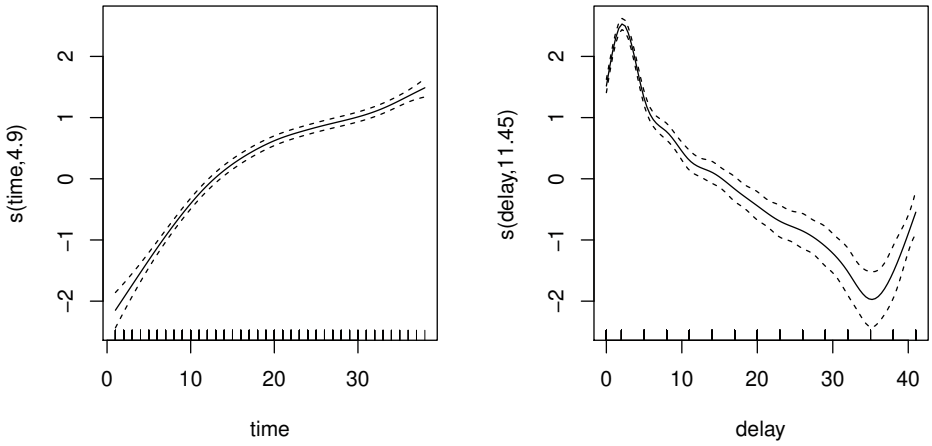
The fits are very similar, but aids.gam2 has slightly lower AIC of 792.0 compared to 792.1 — these are so similar that the choice should be based on interpretability rather than on AIC.

Example: AIDS data



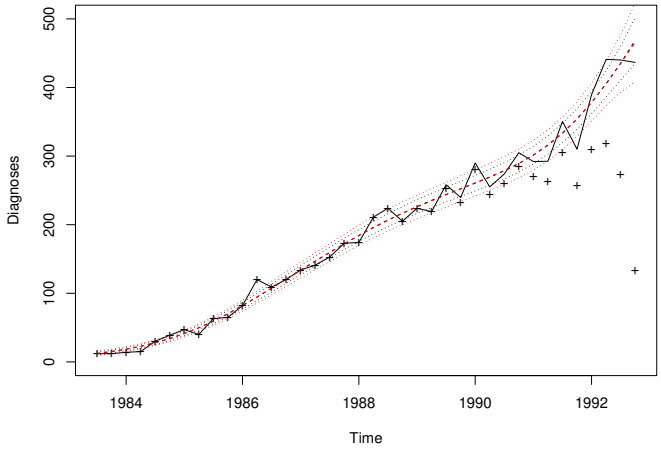
Estimates of (centered) smooth functions $s(j)$ and parameters based on plot(aids.gam1) and coef(aids.gam1).

Example: AIDS data



Estimates of (centered) smooth functions $s(j)$ and $s(k)$ based on plot(aids.gam2).

Example: AIDS data



Numbers of recorded deaths (+), with estimated mean deaths per quarter based on chain-ladder model (solid) and on Poisson (black dashes) and quasi-likelihood GAMs with Poisson variance function $V(\mu) = \mu$ (red dashes). The last two estimates have 95% pointwise confidence intervals (dots) based on the fit (treating the smoothing parameters as fixed). To make these I had to compute the fitted means for the missing lower right triangle of the data table.

Closing

- The basic ideas of regression, dependence of a response on explanatory variables, extend far beyond the linear model, to
  - non-linear dependence on explanatory variables;
  - general response distributions (Poisson, binomial, ...);
  - 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 using basis functions;
  - maximum likelihood inference;
  - estimation using iterative weighted least squares algorithms;
  - penalized fitting to allow for random effects/basis functions;
  - analysis of deviance;
  - residuals and other diagnostics.