Regression Methods

Myrto Limnios

Autumn 2025 - Week 3

The Linear Model - First Results and Geometric Interpretations (*Continued*)

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

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

Lemma 3

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

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

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

Theorem 4 (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 $\sigma^2 (X^{\mathrm{T}} X)^{-1}$.

balance between bias + variance

bias $= 0$

Proof Lemma 3

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

$$\mathrm{E}(\widehat{\beta}) = \mathrm{E}(Ay) = A\mathrm{E}(y) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta = \beta,$$

$$\text{var}(\widehat{\beta}) = A\text{var}(y)A^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}I_n\sigma^2\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}^{\mathrm{T}} = \sigma^2(X^{\mathrm{T}}X)^{-1}.$$

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

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

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

$$\mathrm{E}[\text{tr}\{(I_n-H)yy^{\mathrm{T}}\}] = \text{tr}\{(I_n-H)\mathrm{E}(yy^{\mathrm{T}})\}$$

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

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

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

so $\mathrm{E}(S^2) = \sigma^2$, because

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

Proof Theorem 4

- 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 $n \geq p$).

(Approximate) Sampling Distribution of $\widehat{\beta}$ under Second-order Assumptions

If we only assume $\mathrm{E}[\varepsilon] = 0$ and $\mathrm{var}[\varepsilon] = \sigma^2 I_n$

$\hookrightarrow$then Gauss-Markov says $\widehat{\beta}$ optimal linear unbiased estimator, regardless of distribution of ε.

Question: *What can we say about the sampling distribution of $\widehat{\beta}$ when ε is not necessarily Gaussian?*

Note that we can always write

$$\widehat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon.$$

$y = X\beta + \varepsilon$

$\Leftrightarrow X^T y = X^T X \beta + X^T \varepsilon$

$\Leftrightarrow (X^T X)^{-1} X^T y = \beta + (X^T X)^{-1} X^T \varepsilon$

- Since there is a huge variety of candidate distributions for ε that would be compatible with the property $\mathrm{var}(\varepsilon) = \sigma^2 I_n$, we cannot say very much about the exact distribution of $\widehat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon$.
- Can we at least hope to say something about this distribution asymptotically, as the sample becomes large?

For this, we need an appropriate asymptotic framework for covariates:

- We let $n \to \infty$ (# rows of X tend to infinity)
- # columns of X, i.e., p, (held fixed).

Theorem 5 (Large Sample Distribution of $\widehat{\beta}$)

Let $\{X_n\}_{n\geq 1}$ be a sequence of $n \times p$ design matrices, and $\{\varepsilon_n\}_{n\geq 1}$ a sequence of n-vectors, and define $y_n = X_n\beta + \varepsilon_n$. If

1. *X_n is of full rank p for all $n \geq 1$*
2. *$\max_{1\leq i\leq n}[x_i^\top (X_n^\top X_n)^{-1} x_i] \overset{n\to\infty}{\longrightarrow} 0$, (where $x_i^\top$ is the ith row of X_n)*
3. *$\mathrm{E}[\varepsilon_n] = 0$ and $\mathrm{var}[\varepsilon_n] = \sigma^2 I_n$ for all $n \geq 1$,*

then the least squares estimator $\widehat{\beta}_n = (X_n^\top X_n)^{-1} X_n^\top y_n$ satisfies

$$(X_n^\top X_n)^{1/2}(\widehat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}_p(0, \sigma^2 I_p).$$

Notice that (2) relates to the diagonal elements of the hat matrix $H_n = X_n(X_n^{\mathrm{T}} X_n)^{-1} X_n^{\mathrm{T}}$.

Example

The Theorem states that if its diagonal elements are "well-behaved", then the asymptotic distribution

$$\text{for } n \text{ "large enough", } \widehat{\beta} \overset{d}{\approx} \mathcal{N}\{\beta, \sigma^2 (X_n^\top X_n)^{-1}\}$$

- i.e. distribution of $\widehat{\beta}$ gradually becomes the same as what it would be if ε were Gaussian
- . . . provided design matrix X satisfies extra condition (2).
- Can be shown equivalent to: *diagonal elements of* $H_n = X_n(X_n^\top X_n)^{-1} X_n^\top$ *say* $h_{jj}(n)$ *converge to zero uniformly in* j *as* $n \to \infty$
- Note that $\text{trace}(H) = p$, so that the average $\sum h_{jj}(n)/n \to 0$ — the question is do all the $h_{jj}(n) \to 0$ uniformly?

Has a very clear interpretation in terms of the form of the design that we will see when we discuss the notions of **leverage** and **influence**.

To understand Condition (2), consider simple linear model

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

$X = \begin{pmatrix} 1 & t_1 \\ 1 & \vdots \\ 1 & t_n \end{pmatrix}$ n, $p = 2$

Here, $p = 2$. Can show that

$$h_{jj}(n) = \frac{1}{n} + \frac{(t_j - \bar{t})^2}{\sum_{k=1}^{n}(t_k - \bar{t})^2}$$

H = in terms of X

- Suppose $t_i = i$, for $i = 1, \ldots, n$ (regular grid). Then

$$h_{jj}(n) = \frac{1}{n} + \frac{\{j - (n+1)/2\}^2}{(n^2 - n)/12}$$

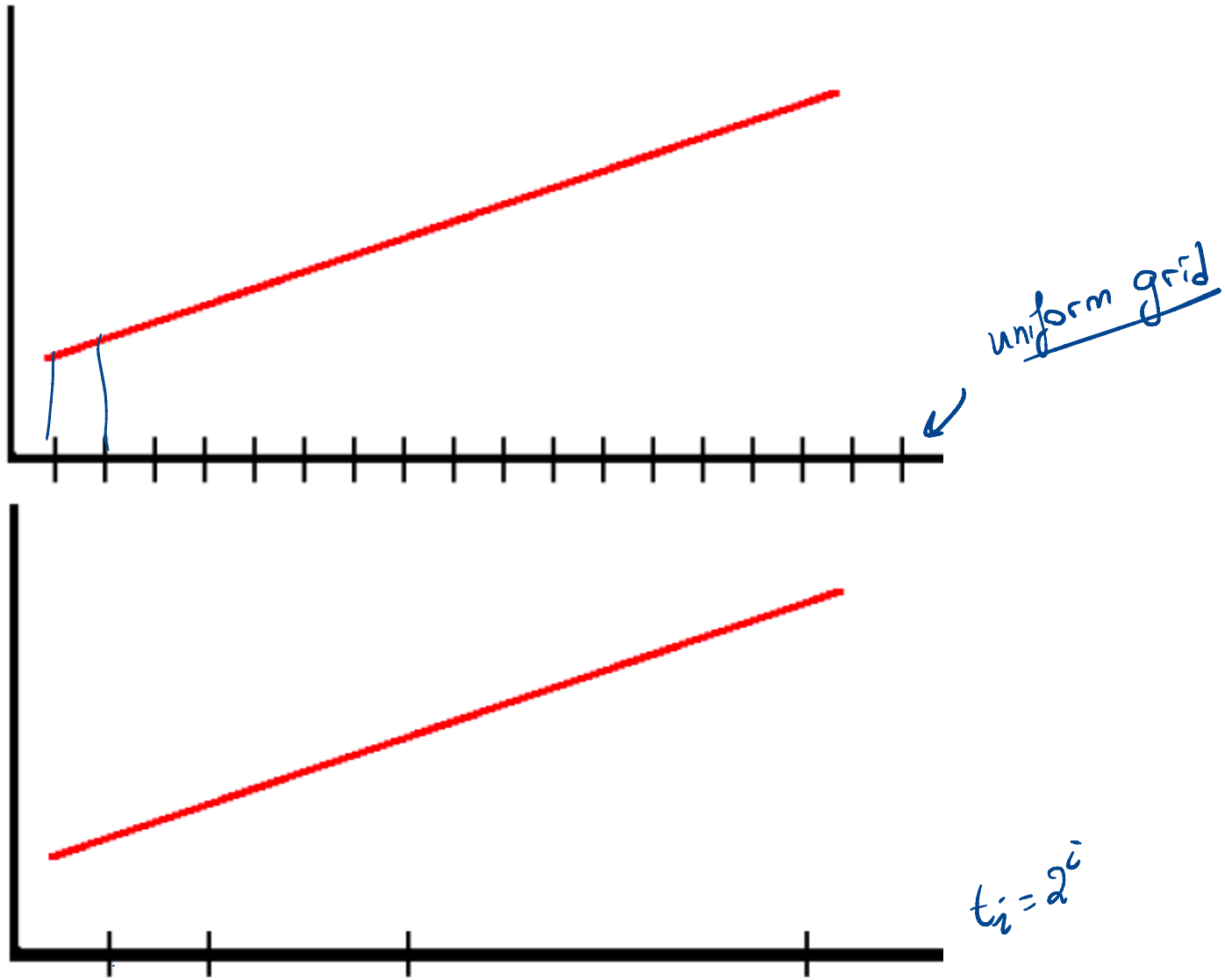
so $$\max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) = \frac{1}{n} + \frac{6(n-1)}{n(n+1)} \overset{n \to \infty}{\longrightarrow} 0.$$

- Now consider $t_i = 2^i$ (grid points spread apart as n grows).
  The centre of mass and sum of squares of the grid points is now

$$\bar{t} = \frac{2(2^n - 1)}{n}, \quad \sum_{i=1}^{n}(t_i - \bar{t})^2 = \frac{4^{n+1} - 4}{3} - \frac{4^{n+1} + 4 - 2^{n+3}}{n}$$

  and so

$$\max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) \overset{n \to \infty}{\longrightarrow} \frac{3}{4}.$$


uniform grid
$t_i = 2^i$

Normal-theory linear model

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

Theorem 6

Under the normal-theory linear model,

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

where $S^2 = \|y - \widehat{y}\|^2/(n-p)$.

Lemma 7

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

Proof Theorem 6

- 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 3.
- 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

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

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

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

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,

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

Proof Lemma 7

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

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

Cochran's Theorem

Recall that we can decompose a design matrix of rank $p \leq n$ as follows: $X = (X_0, X_1, \ldots, X_R)$ and H_r denotes 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 $P_{R+1} = I - H$.

Theorem 8

If $y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$ and a linear model is fitted whose design matrix X is decomposed in $R+1$ column blocs (cf [Lemma 2, week 1]), 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 $\|P_r y\|^2 \sim \sigma^2 \chi^2_{\nu_r - \nu_{r-1}}(\delta_r^2/\sigma^2)$, 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 8 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.

Proof Theorem 8

- As $P_r P_s = 0$ for $r \neq s$, we have $\operatorname{cov}(P_r y, P_s y) = P_r \operatorname{var}(y) P_s^{\mathrm{T}} = \sigma^2 P_r P_s = 0$, 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 7 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 $\nu = \operatorname{rank}(P_r)$. These ranks are $\nu_r - \nu_{r-1}$ for $r = 1, \ldots, R$, and $n - p$ 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

$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$

- Theorem 6 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

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

- 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?)

The Linear Model - Analysis of Variance

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 8 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

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

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

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

ANOVA table

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

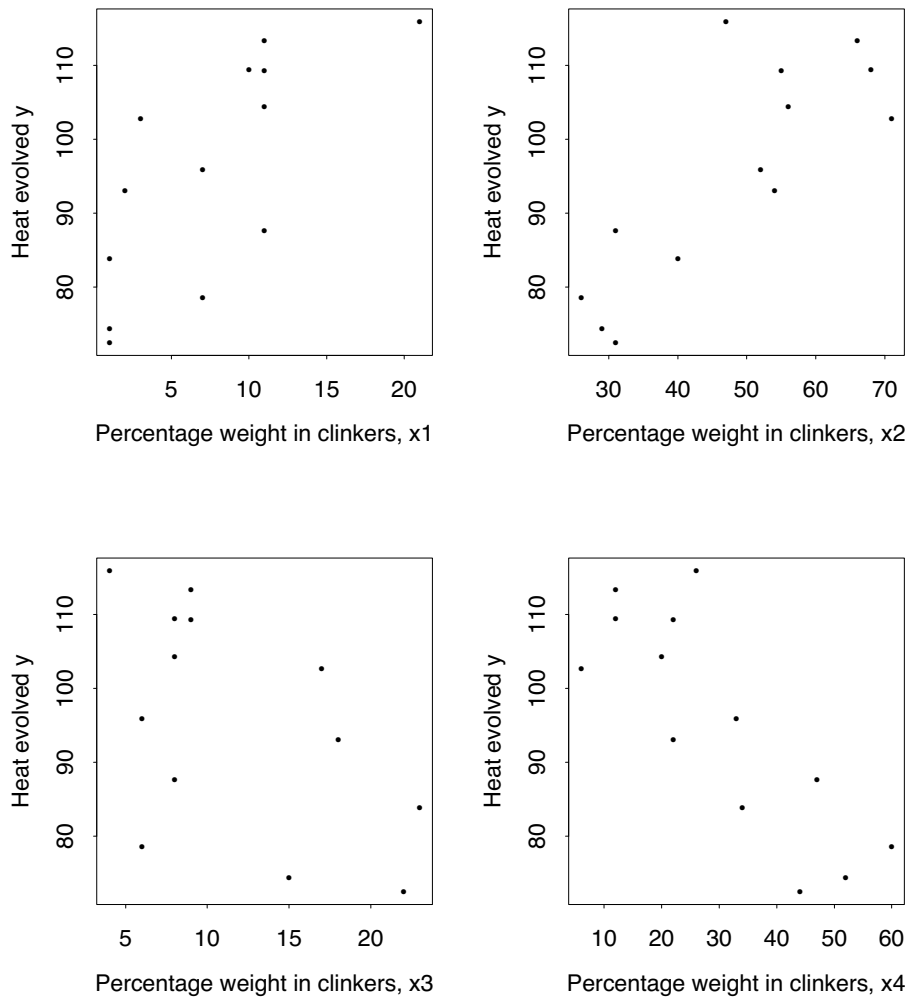
- If $\mu \in \text{span}(X)$ then the residual mean square $\mathbf{MS}_{\text{Res}}$ gives an estimate of σ^2.
- We test for an effect of term X_r by noting that
  - if X_r explains no more than $(X_0, \ldots, X_{r-1})$, then

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

  - if X_r does have additional explanatory power, then the distribution of $\mathbf{MS}_r$ 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.

- 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.

The Linear Model - 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

1. linearity by plotting r_j against the covariates (those in X and those not in X);
2. constant variance by plotting r_j (or $|r_j|$) against fitted values $\widehat{y}_j$;
3. independence by ACF of residuals (if data time-ordered);
4. for outliers, which are visible as unusual residuals; and
5. normality using a normal QQ-plot of r_j.

Checking linearity

A first impression can be drawn by looking at plots of the response against each of the explanatory variables.

Other plots to look at?

Notice that under the assumption of linearity we have

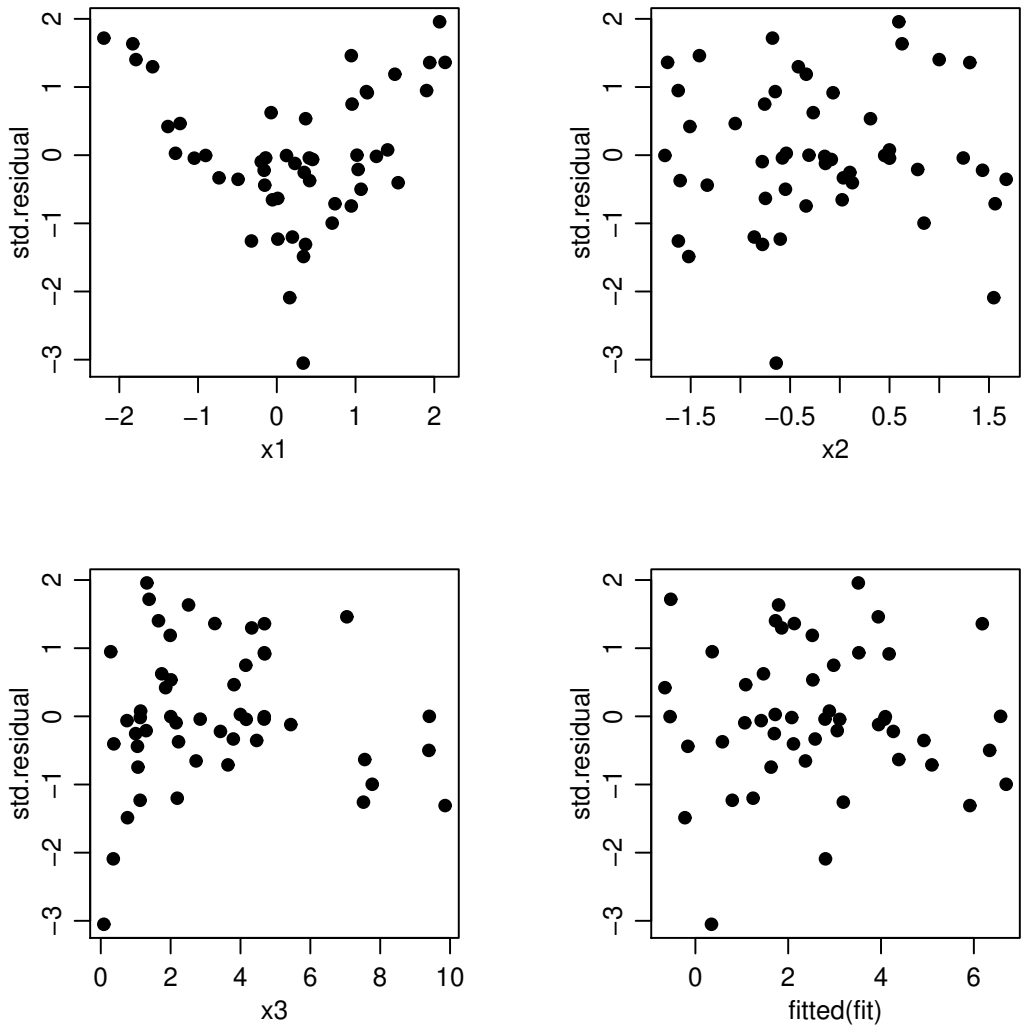
$$X^\top e = 0.$$

Hence, **no correlation should appear between explanatory variables and residuals.**

1. Plot standardised residuals r against each covariate (columns of X).
   - ↪ No systematic patterns should appear in these plots. A systematic pattern would suggest incorrect dependence of the response on the particular explanatory (e.g. need to add a transformation of that explanatory as an additional variable).
2. Plot standardised residuals r against covariates left out of the model.
   - ↪ No systematic patterns should appear in these plots. A systematic pattern suggests that we have left out an explanatory variable that should have been included.

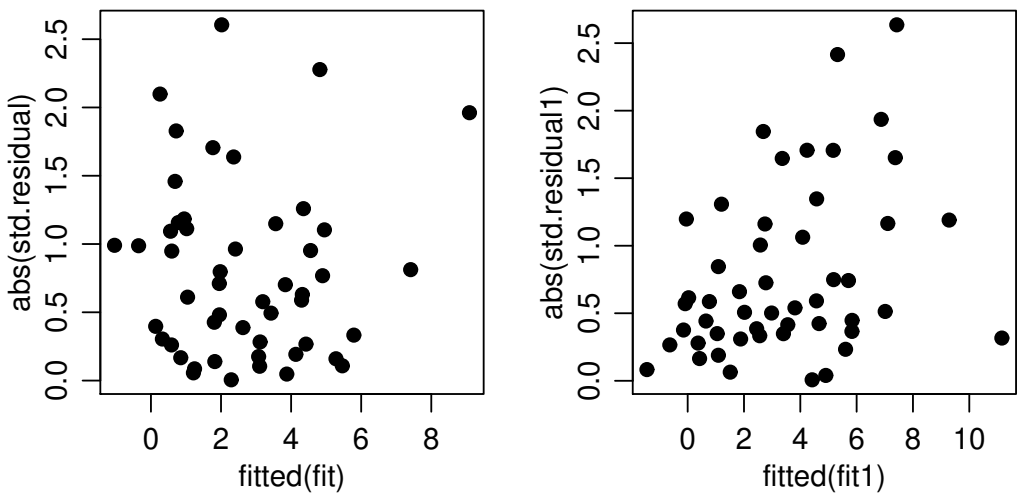
Checking linearity

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



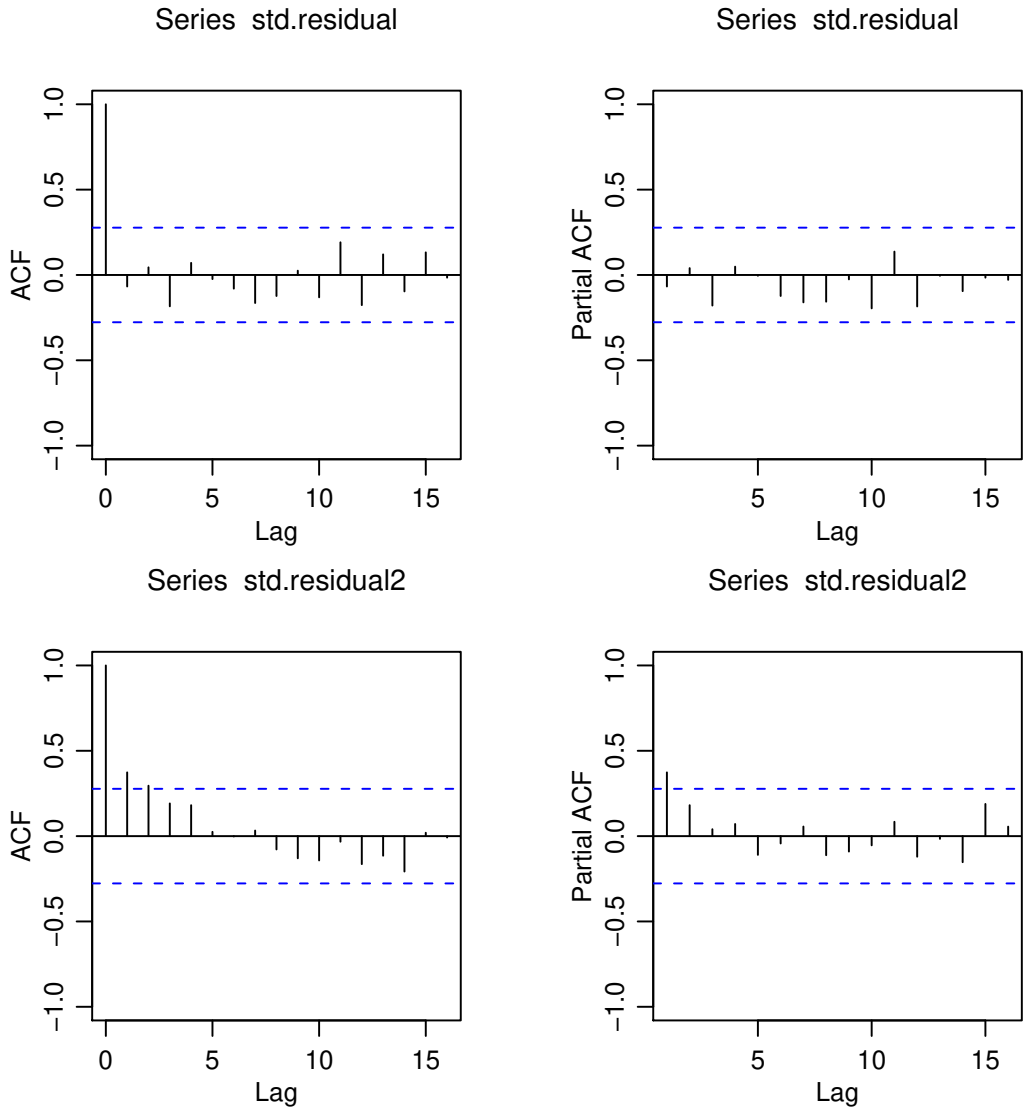
Checking independence

- It is assumed that $\mathrm{var}[\varepsilon] = \sigma^2 \boldsymbol{I}_n$. Under the assumption of normality, it is equivalent to independence.
- Dependence can greatly increase uncertainty of final conclusions.
- Substantive knowledge is helpful in suggesting whether it might be present:
  1. were the data gathered in temporal/spatial/. . . order?
  2. were the data sampled/gathered in groups (e.g., spatial, several observations on different individuals, . . .)?
  3. 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

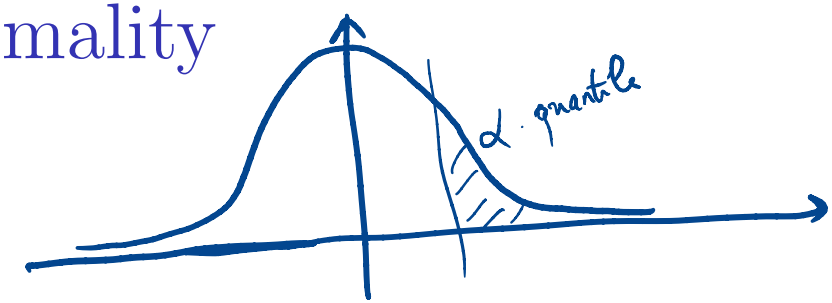
$$\mathrm{corr}(r_j, r_{j+t}), \quad \mathrm{corr}(r_j, r_{j+t} \mid r_{j+1}, \ldots, r_{j+t-1}), \quad t = 1, \ldots$$

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

Checking independence

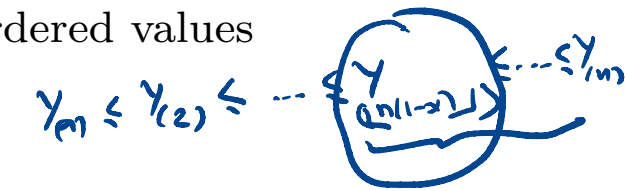


Checking for outliers and normality



- 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.

Reminder: Q-Q plots

Idea: compare the distribution of standardised residuals against a Normal distribution.

How?

Compare **empirical vs theoretical quantiles** ...

Reminer: The α-*quantile* ($\alpha \in [0,1]$) of a distribution F is the value $F^-(\alpha)$ defined as

$$F^-(\alpha) := \inf\{t \in \mathbb{R} : F(t) \geq \alpha\}.$$

Given a sample $W_1, \ldots, W_n$, the *empirical* α *quantile* is the value defined as

$$\widehat{F}^-(\alpha) := \inf\{t \in \mathbb{R} : \widehat{F}(t) \geq \alpha\} = \inf\left\{t \in \mathbb{R} : \frac{\#\{W_i \leq t\}}{n} \geq \alpha\right\}.$$

where $\widehat{F}$ is the empirical distribution function (as defined before).

A **quantile plot** for a given sample plots certain empirical quantiles against the corresponding theoretical quantiles (i.e. those under the assumed distribution).

If the sample at hand originates from F, then we expect that the points of the plot fall close to the 45° line.

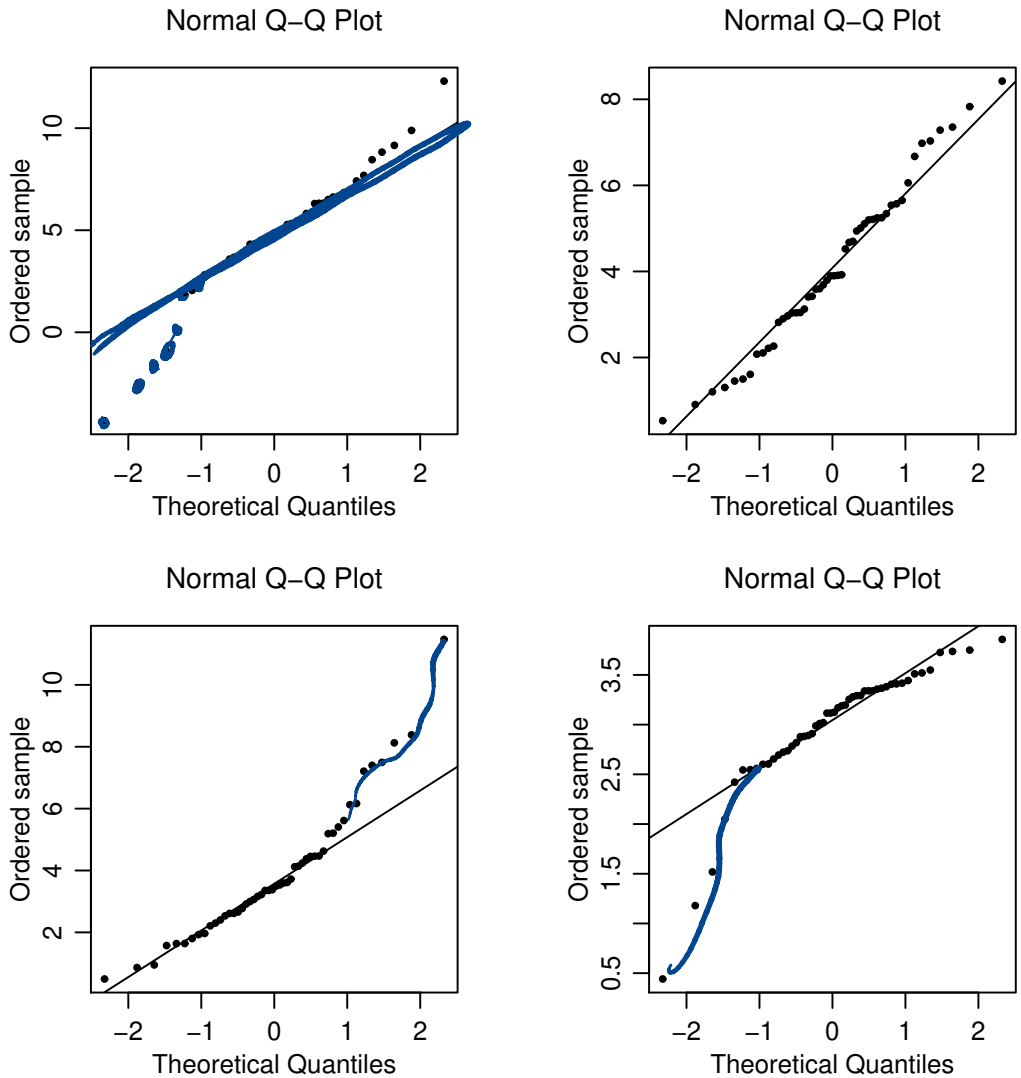
- Plot the empirical $\{k/n\}_{k=1}^{n}$ quantiles of standardised residuals

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

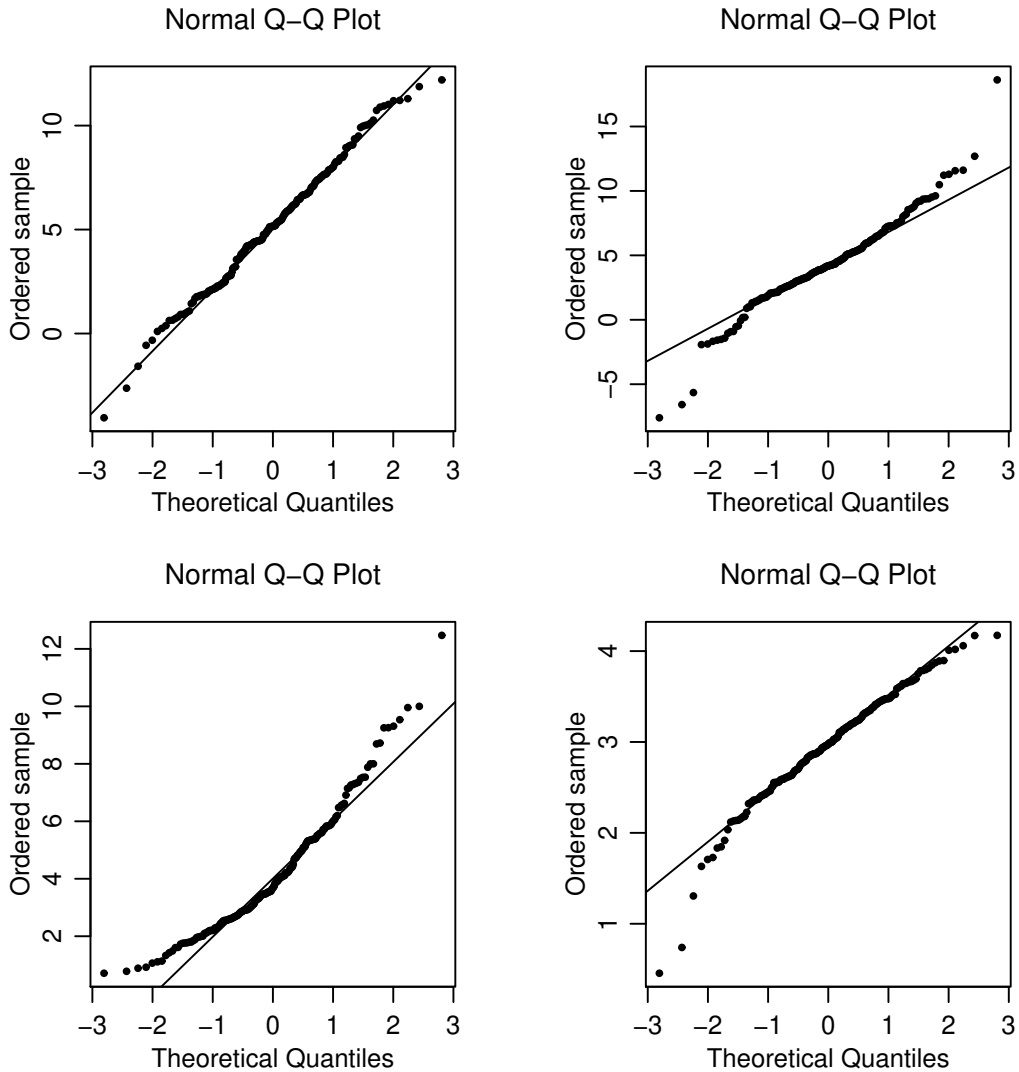
against theoretical quantiles $\Phi^{-1}\{1/(n+1)\}, \ldots, \Phi^{-1}\{n/(n+1)\}$ of a $\mathcal{N}(0,1)$ distribution.

  - ↪ Think why we pick $\Phi^{-1}\left(\frac{k}{n+1}\right)$ instead of $\Phi^{-1}\left(\frac{k}{n}\right)$.
  - ↪ If the points of the quantile plot deviate significantly from the 45° line, there is evidence against the normality assumption. Outliers, skewness and heavy tails easily revealed.
  - ↪ If we plot the empirical quantiles of the unstandardised residuals against those of a $N(0,1)$, then we compare against a line with slope equal to stdev(e) and intercept zero.

Checking normality, $n = 50$



Checking normality, $n = 200$



Leverage and influence

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

Response transformation

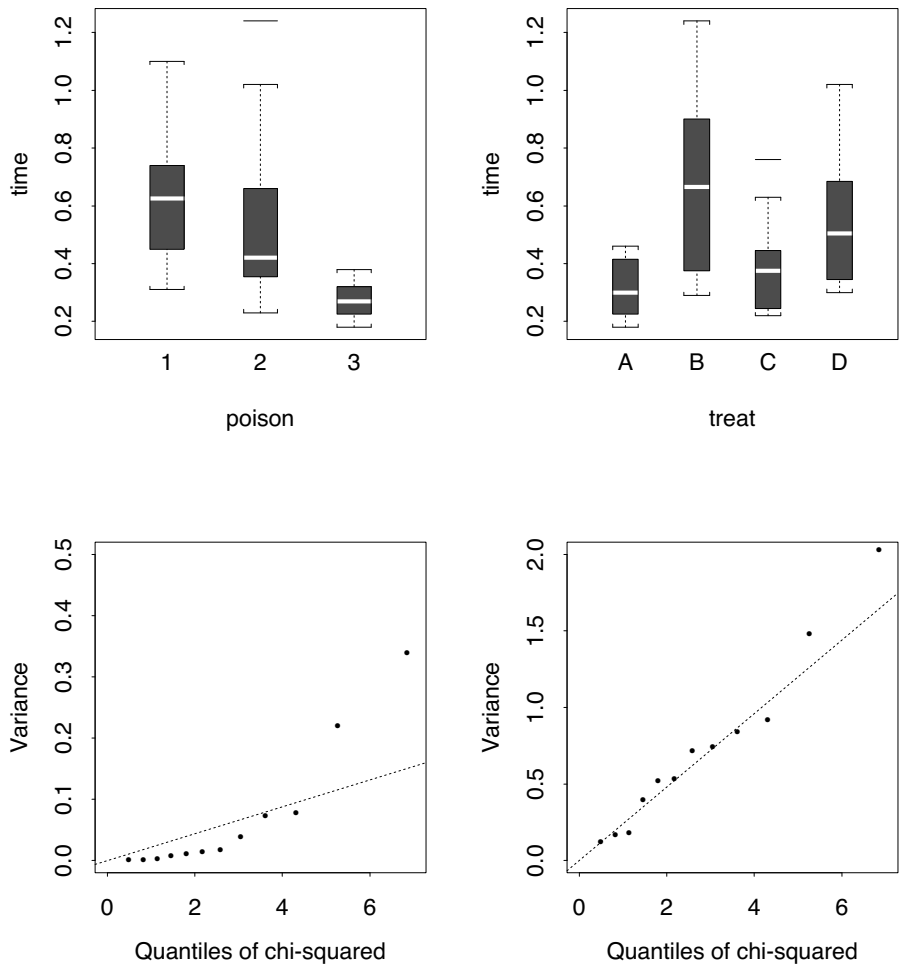
- Linear model for y may be better applied for some transformation $g(y)$, especially if some y are much larger than others, or the variance is non-constant.
- Survival times y_{ptj} in 10-hour units of animals in a 3×4 factorial experiment with four replicates, with (below) average (standard deviation)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

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

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

Example: Poison data

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



Box–Cox transformation

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

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

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

- Suppose normal linear model

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

  applies for some β, σ and λ to be determined. Here X contains 1_n, so use of $y^{(\lambda)}$ just changes intercept and rescales β and σ.

- Use profile log likelihood for λ to choose 'best' transformation (usually from list above to aid interpretation).

- Interpretation of β depends on λ, so usually we ignore the fact that λ was estimated, unless we are not interested in β (e.g., when performing 'automatic' prediction).

Example: Poison data

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

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

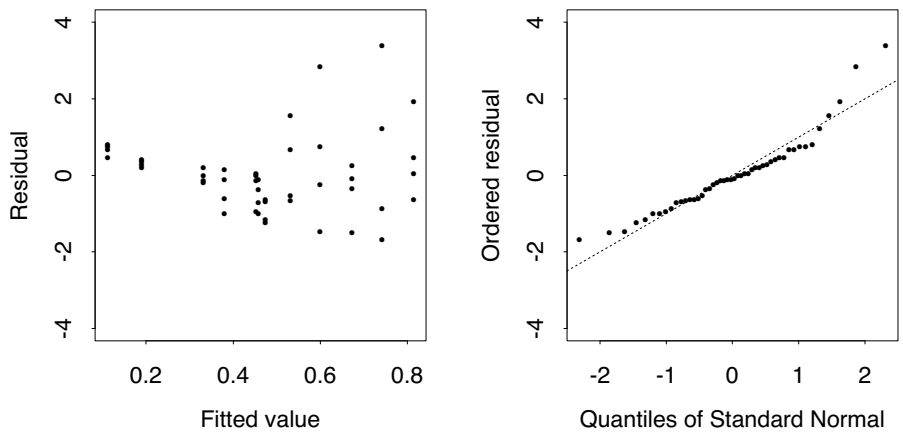
- Analyses of variance with responses y and y^{-1}. For MS and F read 'Mean square' and 'F statistic'.
- The terms explain appreciably more of the variation of y^{-1}, suggesting that this is a preferable choice of response.

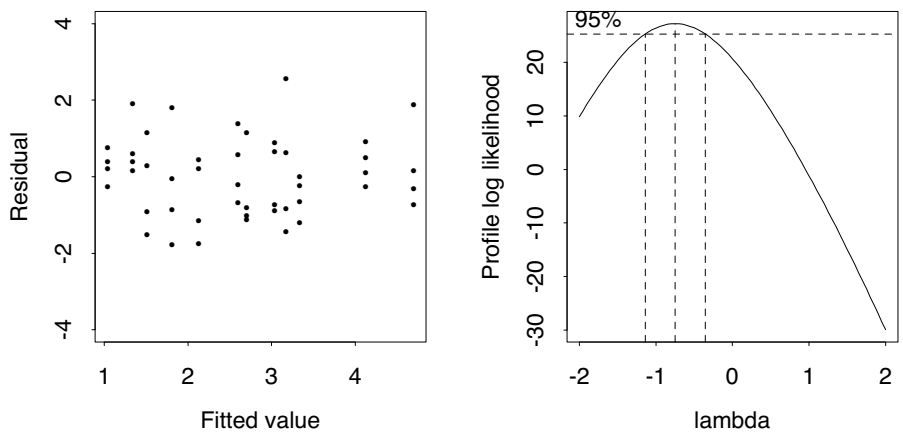
Term	df	Response y			Response y^{-1}		
		SS	MS	F	SS	MS	F
Poisons	2	1.033	0.517	23.22	34.88	17.44	72.63
Treatments	3	0.921	0.307	13.81	20.41	6.80	28.34
Treatments × Poisons	6	0.250	0.042	1.87	1.57	0.26	1.09
Residual	36	0.801	0.022		8.64	0.24	

$y = \mu + \alpha_t + \beta_p + \varepsilon$ $\quad$ $y^{-1} = \mu' + \alpha'_t + \beta'_p + \varepsilon'$

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!

The Linear Model - 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).