Linear Models

Victor Panaretos

Institut de Mathématiques – EPFL

victor.panaretos@epfl.ch

Statistical model for:

- Y (random variable) $\overset{\text{depending on}}{\longleftarrow}$ x (non-random variable)

Aim: understand the effect of x on the random quantity Y

General formulation[1]:

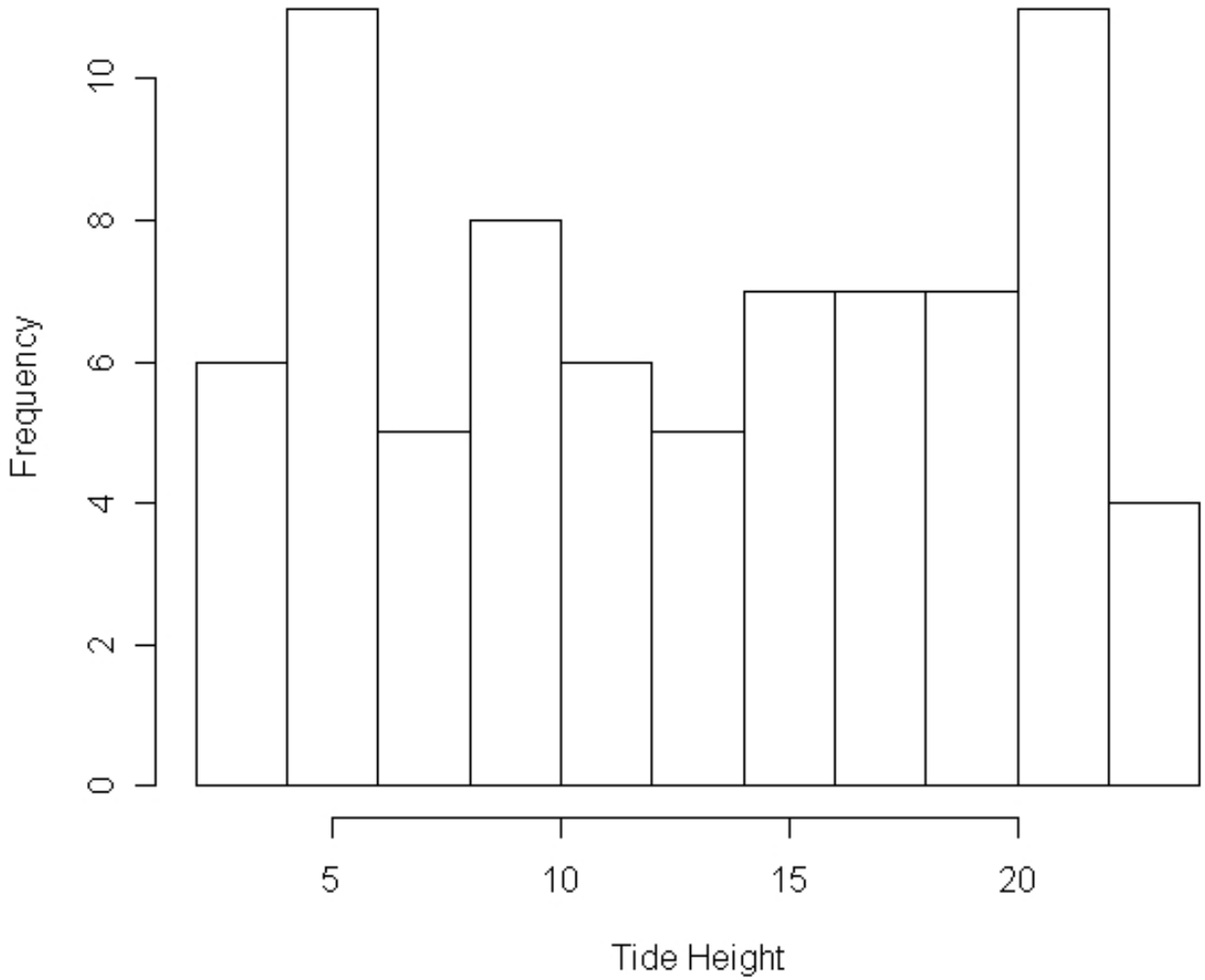
$$Y \sim \text{Distribution}\{g(x)\}$$

Statistical Problem: Estimate (learn) $g(\cdot)$ from data $\{(x_i, y_i)\}_{i=1}^n$. Use for:

- Description
- Inference
- Prediction
- Data compression (parsimonious representations)
- ...

[1] Often books/people write $Y \mid x \sim \text{Distribution}\{g(x)\}$ but this implies that (X, Y) have a joint distribution; this assumption is unnecessary (e.g., in a designed experiment we choose values for x). Despite this, we write $Y \mid x$ to remind ourselves that the distribution of Y depends on x.

Example: Honolulu tide



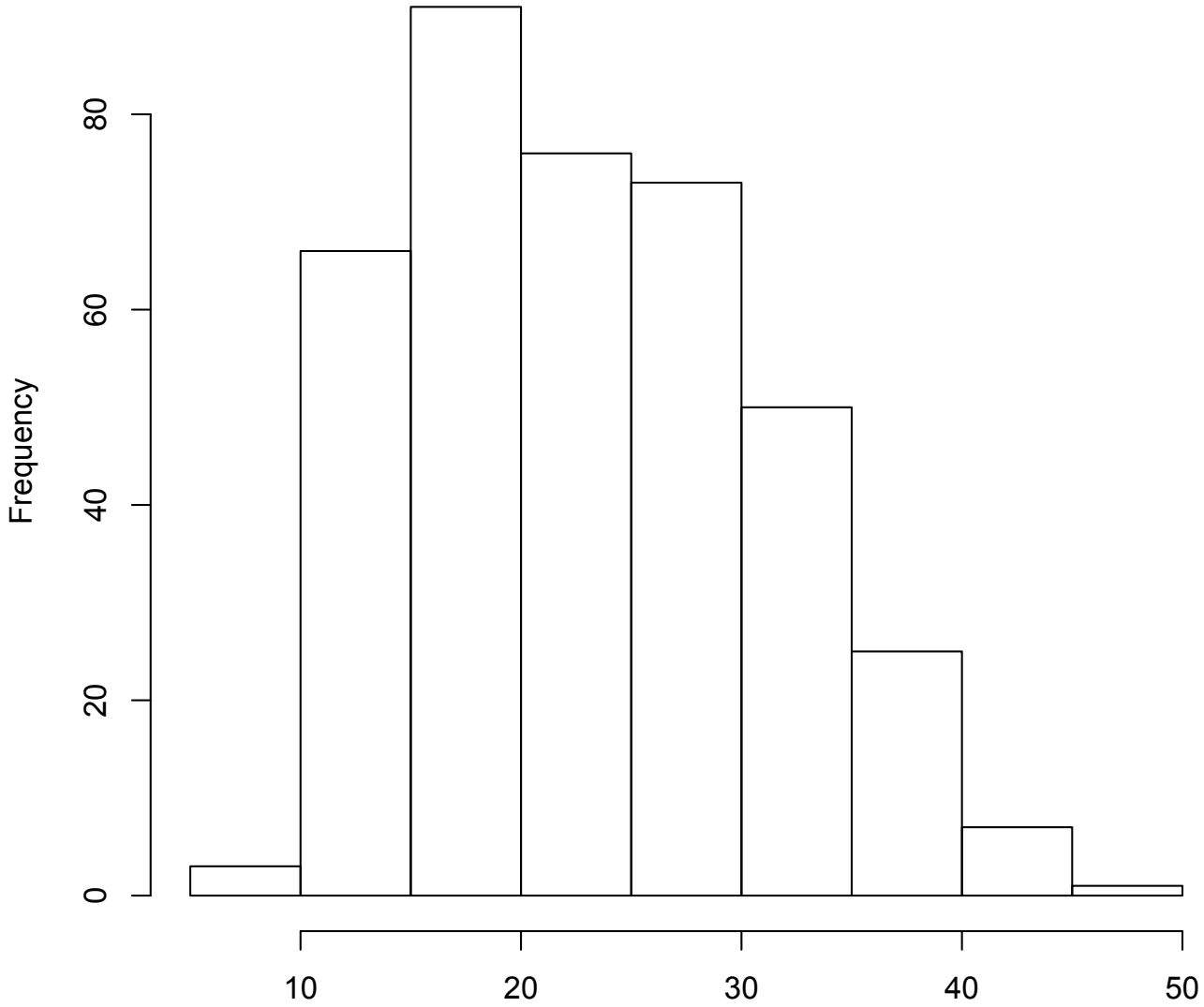
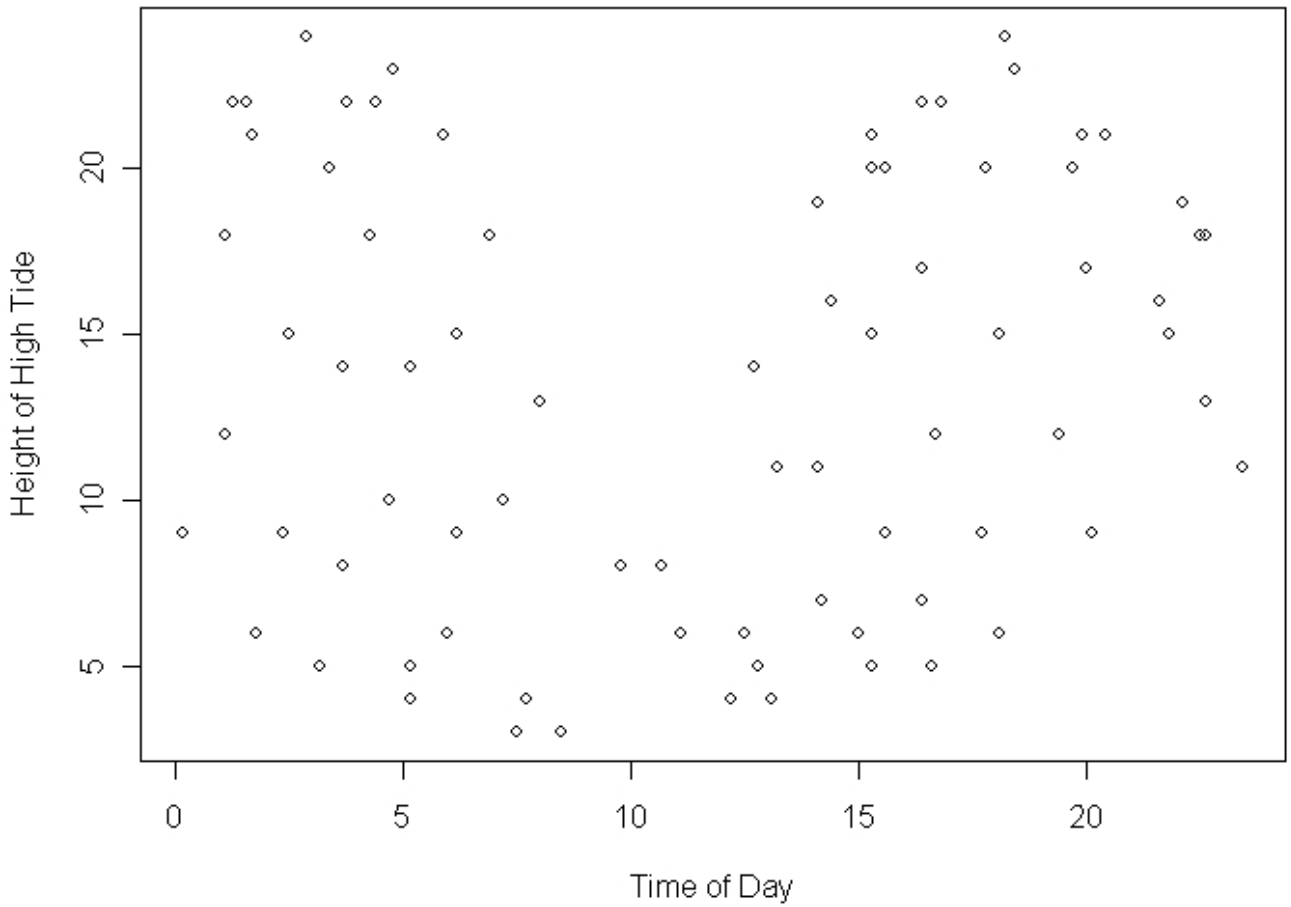
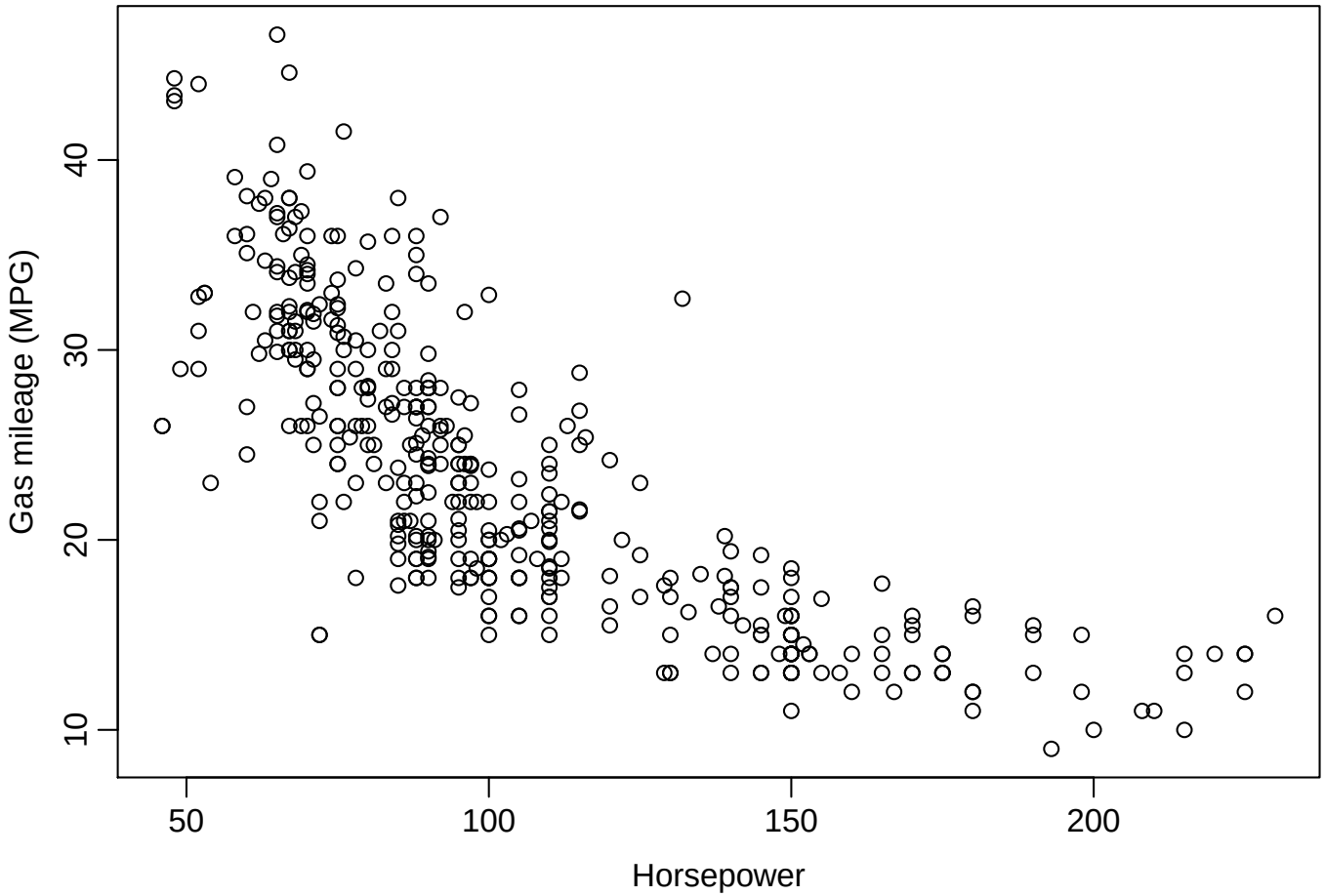


Figure: Miles per gallon for 392 car models

Example: Honolulu tide with time covariate

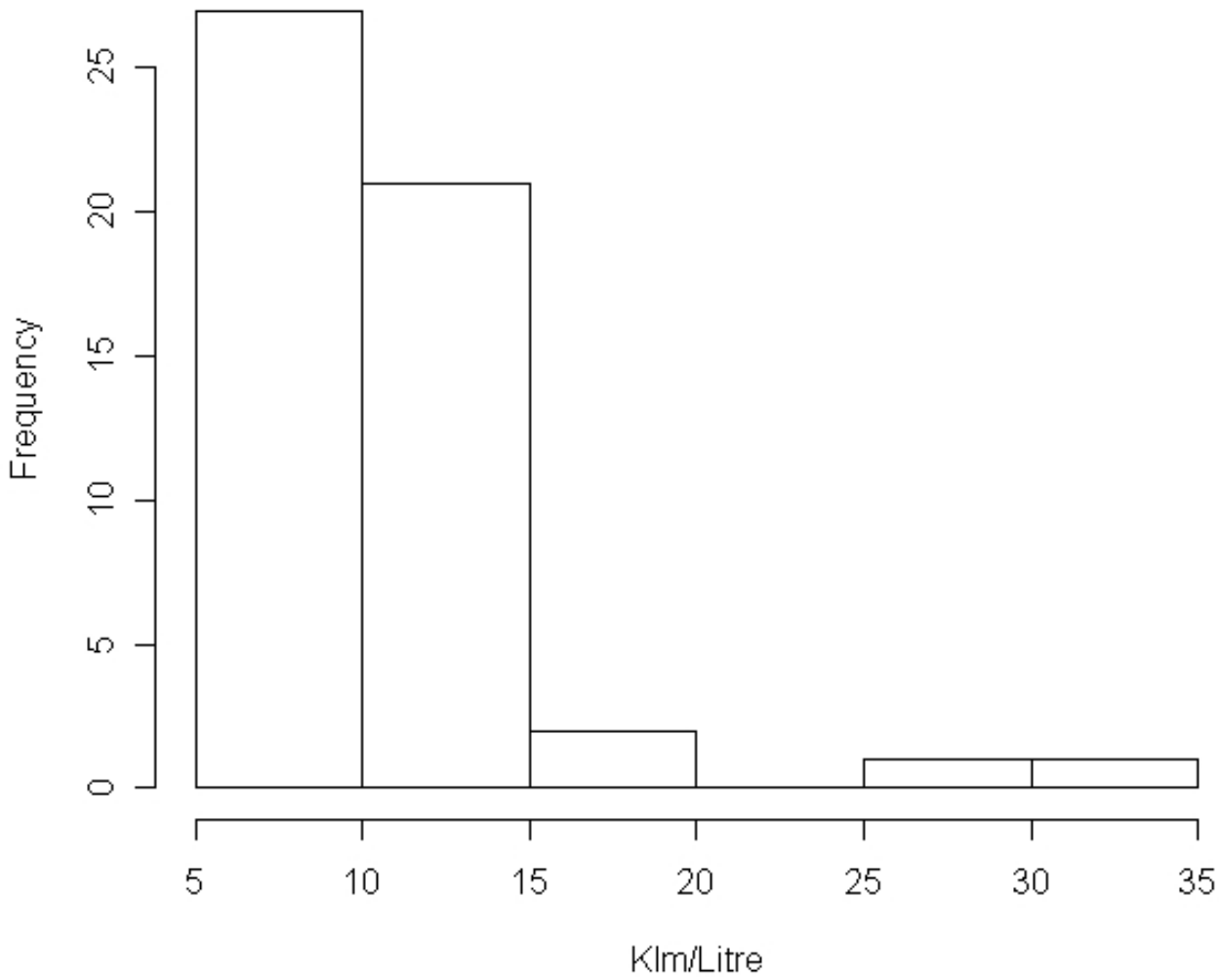


Example: Gass mileage with horsepower covariate

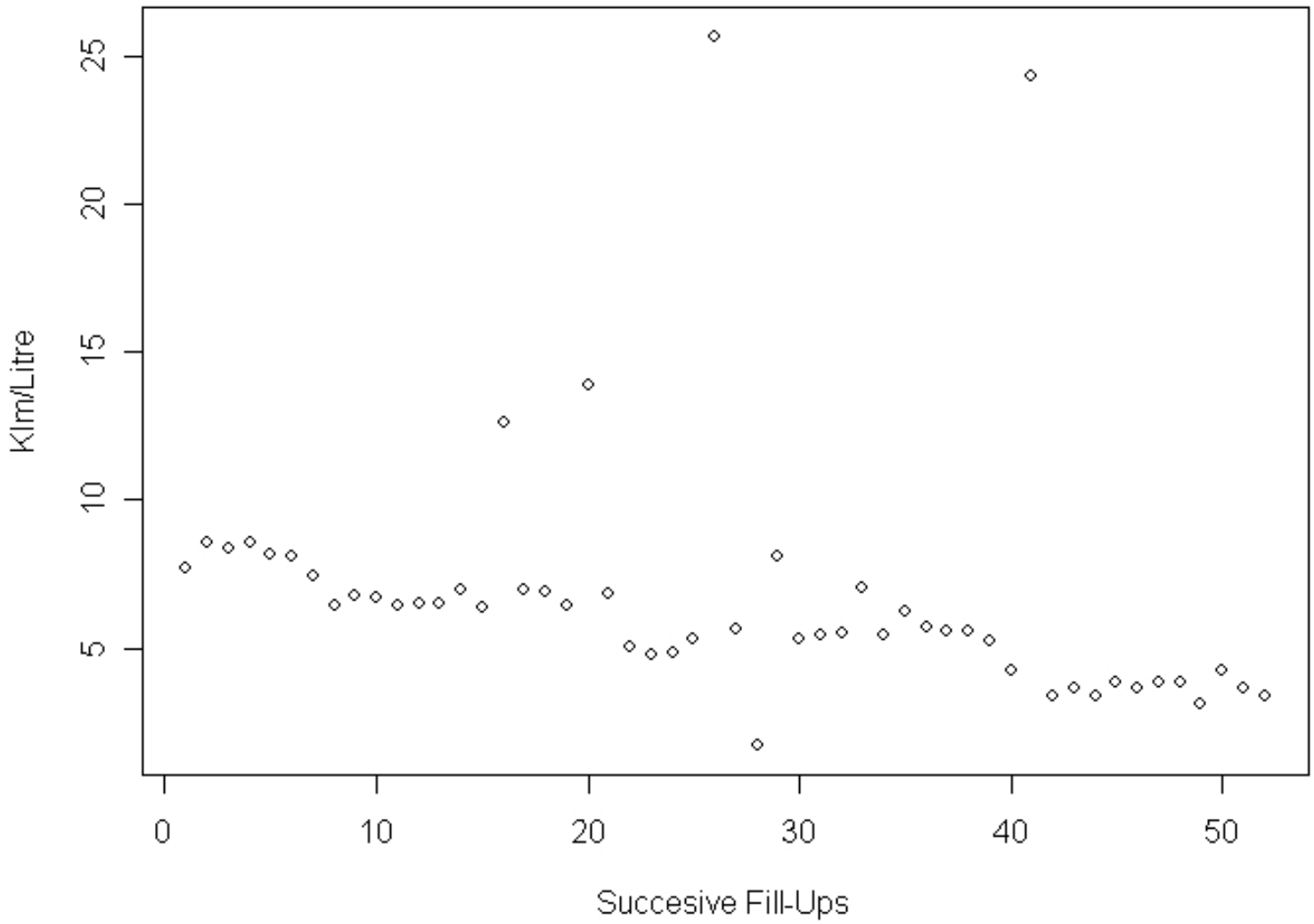


Example: Professor's Van

Example: Professor's Van



Example: Professor's Van



Remember general model:

$$Y \sim \text{Distribution}\{g(x)\}$$

x can be:

- continuous, discrete, categorical, vector ...
- arrive randomly, or be chosen by experimenter, or both
- however x arises, we treat it as constant in the analysis

Distribution can be:

- Gaussian (Normal), Laplace, binomial, Poisson, gamma, General exponential family, ...

Function $g(\cdot)$ can be:

- $g(x) = \beta_0 + \beta_1 x$, $g(x) = \sum_{k=-K}^{K} \beta_k e^{-ikx}$, Cubic spline, ...

- $Y, x \in \mathbb{R}$, $g(x) = \beta_0 + \beta_1 x$, Distribution $=$ Gaussian

$$Y \mid x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

$$\Updownarrow$$

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

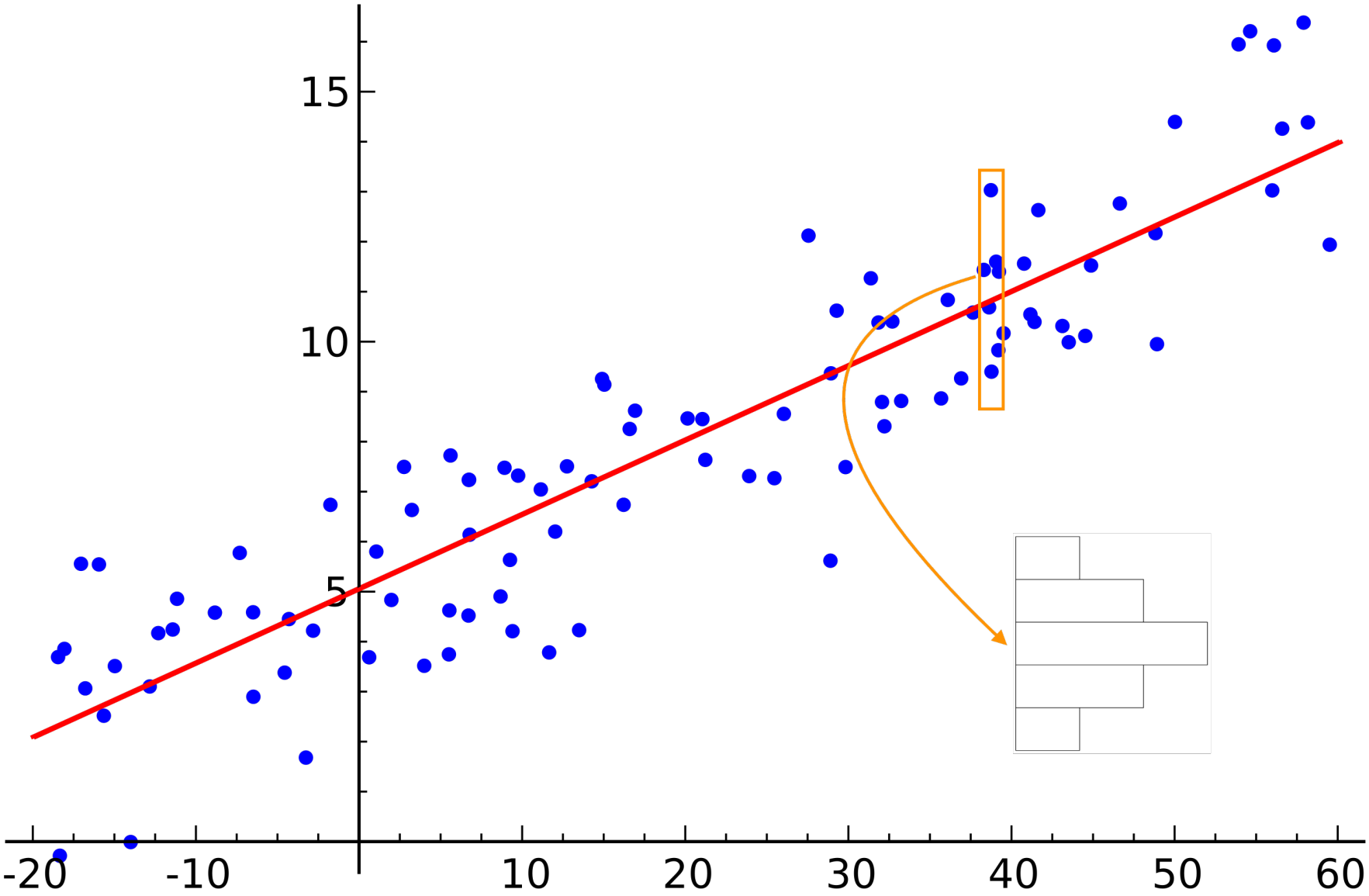
The second verson is useful for mathematical work, but is puzzling statistically, since we don't observe ϵ.

- Also, x could be vector ($Y, \beta_0 \in \mathbb{R}$, $x \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$):

$$Y \mid x \sim \mathcal{N}(\beta_0 + \beta^\top x, \sigma^2)$$

$$\Updownarrow$$

$$Y = \beta_0 + \beta^\top x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$


15
10
5
-20
-10
10
20
30
40
50
60

Start from Normal linear model $\longrightarrow$ gradually generalise . . .
Important features of Normal linear model:

- Gaussian distribution
- Linearity

These two combine well and give geometric insights to solve the estimation problem. Thus we need to revise some linear algebra and probability . . .

Will base course on the Gaussian assumption, but relax linearity later:

- linear Gaussian regression
- nonlinear Gaussian regression
- nonparametric Gaussian regression

Many further generalisations are possible . . .

Projections, Spectra, Gaussian Law

If Q is an $n \times p$ real matrix, we define the *column space* (or *range*) of Q to be the set spanned by its columns:

$$\mathcal{M}(Q) = \{y \in \mathbb{R}^n : \exists \beta \in \mathbb{R}^p, \ y = Q\beta\}.$$

- Recall that $\mathcal{M}(Q)$ is a subspace of $\mathbb{R}^n$.
- The columns of Q provide a coordinate system for the subspace $\mathcal{M}(Q)$
- If Q is of full column rank (p), then the coordinates β corresponding to a $y \in \mathcal{M}(Q)$ are unique.
- Allows interpretation of system of linear equations

$$Q\beta = y.$$

[existence of solution $\leftrightarrow$ is y an element of $\mathcal{M}(Q)$?]
[uniqueness of solution $\leftrightarrow$ is there a unique coordinate vector β?]

Two further important subspaces associated with a real $n \times p$ matrix Q:

- the *null space* (or *kernel*), $\ker(Q)$, of Q is the subspace defined as

$$\ker(Q) = \{x \in \mathbb{R}^p : Qx = 0\};$$

- the *orthogonal complement* of $\mathcal{M}(Q)$, $\mathcal{M}^{\perp}(Q)$, is the subspace defined as

$$\begin{aligned} \mathcal{M}^{\perp}(Q) &= \{y \in \mathbb{R}^n : y^{\top} Qx = 0,\ \forall x \in \mathbb{R}^p\} \\ &= \{y \in \mathbb{R}^n : y^{\top} v = 0,\ \forall v \in \mathcal{M}(Q)\}. \end{aligned}$$

The orthogonal complement may be defined for arbitrary subspaces by using the second equality.

Theorem (Singular Value Decomposition)

Any $n \times p$ real matrix can be factorised as

$$\underset{n\times p}{Q} = \underset{n\times n}{U}\ \underset{n\times p}{\Sigma}\ \underset{p\times p}{V}^{\top},$$

where U and $V^{\top}$ are orthogonal with columns called left singular vectors and right singular vectors, respectively, and Σ is diagonal with non-negative real entries called singular values.

1. The left singular vectors corresponding to non-zero singular values form an orthonormal basis for $\mathcal{M}(Q)$.
2. The left singular vectors corresponding to zero singular values form an orthonormal basis for $\mathcal{M}^{\perp}(Q)$.
3. Writing $\{u_i\}_{i=1}^n$ for the left singular vectors and $\{v_j\}_{i=1}^n$ for the right singular vectors, the SVD can also be expressed as
$$\underset{n\times p}{Q} = \sum_{j=1}^{\mathtt{rank}(Q)} \sigma_j \underbrace{u_j}_{n\times 1} \underbrace{v_j^{\top}}_{1\times p}.$$
4. Obviously, if Q has SVD $Q = U\Sigma V^{\top}$, then $Q^{\top}$ has SVD $Q^{\top} = V\Sigma U^{\top}$

Proof.

Since the statement is invariant to transposition, assume wlog that $n \geq p$. We will prove the statement by induction on p. Assume that $p = 1$ so that Q is a column vector. Then the statement holds true trivially, by taking

$$V^\top = V = 1, \quad \Sigma = \left(\|Q\|, \mathbf{0}_{1\times(n-1)}\right)^\top \quad U = (u_1 \dots u_n), \quad u_1 = Q/\|Q\|$$

and $(u_2, ..., u_n)$ an orthonormal basis for $\text{span}^\perp(u_1)$. Thus the statement is true for all $n \geq p$ when $p = 1$. This is the base case for our induction. For the inductive step, assume that the statement is true for some $p > 1$ and all $n \geq p$. Let us prove that it is also true for $p + 1$ and all $n \geq p + 1$.

Let $\mathbb{S}^{p+1} = \{x \in \mathbb{R}^{p+1} : \|x\| = 1\}$ and $q(x) = \|Qx\|$. Since $q(\cdot)$ is continuous and $\mathbb{S}^{p+1}$ is compact, we have that $q(x)$ is bounded over $\mathbb{S}^{p+1}$ and attains its bounds. So there exists $v_1 \in \mathbb{S}^{p+1}$ such that

$$q(v_1) = \max_{x\in\mathbb{S}^{p+1}} q(x) = \sigma_1 < \infty.$$

Define $u_1 = \sigma_1^{-1} Q v_1$ so $\|u_1\| = 1$. Given any orthonormal bases $\{u_j\}_{j=2}^n$ for $\text{span}^\perp(u_1)$ and $\{v_j\}_{j=2}^p$ for $\text{span}^\perp(v_1)$ define U and V to be orthogonal matrices

$$U = (u_1 \ u_2 \ \dots \ u_n) = (u_1 \ U_1) \qquad \& \qquad V = (v_1 \ v_2 \ \dots \ v_n) = (v_1 \ V_1).$$

Using block matrix multiplication, we see that

$$\underset{n\times n}{U^\top}\ \underset{n\times(p+1)}{Q}\ \underset{(p+1)\times(p+1)}{V} = \begin{pmatrix} u_1^\top \\ U_1^\top \end{pmatrix} Q \begin{pmatrix} v_1 & V_1 \end{pmatrix} = \begin{pmatrix} u_1^\top Q v_1 & u_1^\top Q V_1 \\ U_1^\top Q v_1 & U_1^\top Q V_1 \end{pmatrix}$$

$$= \begin{pmatrix} \underset{1\times 1}{\sigma_1} & \underset{1\times p}{\theta^\top} \\ \underset{(n-1)\times 1}{\mathbf{0}} & \underset{(n-1)\times p}{Z} \end{pmatrix}.$$

Now we claim that $\theta = 0$. To see this, first observe that

$$\sigma_1 = \max_{x\in\mathbb{S}^{p+1}} \|Qx\| = \max_{x\in\mathbb{S}^{p+1}} \|U^\top Qx\| = \max_{x\in\mathbb{S}^{p+1}} \|U^\top QVx\|.$$

Next, let's consider the norm of $U^\top QV \begin{pmatrix} \sigma_1 \\ \theta \end{pmatrix}$,

$$\left\| \begin{pmatrix} \sigma_1 & \theta^\top \\ \mathbf{0} & Z \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \theta \end{pmatrix} \right\| = \left\| \begin{pmatrix} \sigma_1^2 + \theta^\top\theta \\ Z\theta \end{pmatrix} \right\| = \sqrt{(\sigma_1^2 + \theta^\top\theta)^2 + \|Z\theta\|^2}$$

$$\geq \sigma_1^2 + \theta^\top\theta = (\sigma_1^2 + \theta^\top\theta)^{1/2} \left\| \begin{pmatrix} \sigma_1 \\ \theta \end{pmatrix} \right\|.$$

Dividing across by $\|(\sigma_1 \ \theta)^\top\|$, we see that we must necessarily have

$$(\sigma_1^2 + \theta^\top\theta)^{1/2} \leq \max_{x\in\mathbb{S}^{p+1}} \|U^\top QVx\| = \sigma_1 = (\sigma_1^2 + 0)^{1/2}.$$

and so it must be that $\theta^\top\theta = 0$. We conclude that

$$U^\top QV = \begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & Z \end{pmatrix} \overset{thus}{\Longrightarrow} Q = U \begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & Z \end{pmatrix} V^\top.$$

But Z is an $(n-1)\times p$ matrix, and since $n \geq p+1$ it holds that $n-1 \geq p$. So

$$Z_{(n-1)\times p} = W_{(n-1)\times(n-1)}\Omega_{(n-1)\times p}R^\top_{p\times p}$$

where W, R are orthogonal and Ω is diagonal, by our inductive hypothesis. Thus

$$Q_{n\times p} = U_{n\times n}\begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & W\Omega R^\top \end{pmatrix} V^\top_{p\times p} =$$

$$= \underbrace{U\begin{pmatrix} 1 & \mathbf{0}_{1\times(n-1)} \\ \mathbf{0}_{(n-1)\times 1} & W_{(n-1)\times(n-1)} \end{pmatrix}}_{\text{orthogonal}} \underbrace{\begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & \Omega_{(n-1)\times p} \end{pmatrix}}_{\text{diagonal}} \underbrace{\begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & R^\top_{p\times p} \end{pmatrix} V^\top}_{\text{orthogonal}}$$

□

Theorem (Spectral Theorem)

A $p \times p$ matrix A is symmetric if and only if there exists a $p \times p$ orthogonal matrix U and a real diagonal matrix Λ such that

$$A = U \Lambda U^\top.$$

In particular:

1. the columns of $U = (u_1 \ \cdots \ u_p)$ are **eigenvectors** of A, i.e.
$$A u_j = \lambda_j u_j, \qquad j = 1, \ldots, p$$
where $\text{diag}(\lambda_1, \ldots, \lambda_p) = \Lambda$ are the corresponding (real) **eigenvalues** of A.
2. the rank of A is the number of non-zero eigenvalues.
3. if the eigenvalues are distinct, the eigenvectors are unique (up to re-ordering and sign flips).
4. The spectral representation can also be expressed as
$$\underset{p \times p}{A} = \sum_{j=1}^{\texttt{rank}(A)} \lambda_j \underbrace{u_j}_{p \times 1} \underbrace{u_j^\top}_{1 \times p}.$$

Proof.

If $A = 0$, the statement holds trivially, so let $A = A^\top \neq 0$.

First note that the SVDs of $A = U\Sigma V^\top$ and $A^\top = V\Sigma U^\top$ guarantee the existence of a singular vector pair (u, v) with non-zero singular value σ, such that $Av = \sigma u$ and $A^\top u^\top = \sigma v$ so that

$$A(v+u) = Av + Au \overset{\text{by symmetry}}{=} Av + A^\top u = \sigma u + \sigma v = \sigma(u+v).$$

hence $w = u + v$ is an eigenvector of A with real eigenvalue σ.
Now the theorem is obviously true for 1×1 matrices (scalars). So use induction.
Assume any non-zero $p \times p$ symmetric matrix satisfies the theorem statement.
Let $A = A^\top \neq 0$ be $(p+1) \times (p+1)$. By (1), A has at least one eigenvector $w \in \mathbb{R}^p$ with real eigenvalue $\sigma \neq 0$.
Let $W = (w \;\; R)$ where R has p orthonormal columns spanning $\text{span}^\perp(w)$. Then

$$W^\top A W = \begin{pmatrix} w^\top \\ R^\top \end{pmatrix} A \begin{pmatrix} w & R \end{pmatrix} = \begin{pmatrix} w^\top Aw & w^\top AR \\ R^\top Aw & R^\top AR \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & (Aw)^\top R \\ R^\top Aw & R^\top AR \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & R^\top AR \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & B \end{pmatrix}$$

where $B = R^\top AR$ is a symmetric $p \times p$ matrix.

Since B is symmetric, we have $B = V \Omega V^\top$ for $V_{p\times p}$ orthogonal and $\Omega_{p\times p}$ diagonal by our induction hypothesis. In summary

$$A = W \begin{pmatrix} \sigma^2 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & B \end{pmatrix} W^\top$$

$$= \underbrace{W \begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & V_{p\times p} \end{pmatrix}}_{\text{orthogonal}} \underbrace{\begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \Omega_{p\times p} \end{pmatrix}}_{\text{diagonal}} \underbrace{\begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & V^\top_{p\times p} \end{pmatrix} W^\top_{p\times p}}_{\text{orthogonal}}$$

$$= U \Lambda U^\top$$

□

Combining the SVD and the spectral theorem, we notice that:

1. The left singular vectors of Q are eigenvectors of $A = QQ^\top$.
2. The right singular vectors of Q are eigenvectors of $A = Q^\top Q$.
3. The squared singular values of Q are eigenvalues of both $QQ^\top$ and $Q^\top Q$.

A matrix Q is called *idempotent* if $Q^2 = Q$.

An *orthogonal projection* (henceforth projection) onto a subspace $\mathcal{V}$ is a symmetric idempotent matrix H such that $\mathcal{M}(H) = \mathcal{V}$.

Proposition

The only possible eigenvalues of a projection matrix are 0 and 1.

Proposition

Let $\mathcal{V}$ be a subspace and H be a projection onto $\mathcal{V}$. Then $I - H$ is the projection matrix onto $\mathcal{V}^\perp$.

Proof.

$(I - H)^\top = I - H^\top = I - H$ since H is symmetric and,
$(I - H)^2 = I^2 - 2H + H^2 = I - H$. Thus $I - H$ is a projection matrix.

It remains to identify the column space of $I - H$. Let $H = U\Lambda U^\top$ be the spectral decomposition of H. Then $I - H = UU^\top - U\Lambda U^\top = U(I - \Lambda)U^\top$. Hence the column space of $I - H$ is spanned by the eigenvectors of H corresponding to zero eigenvalues of H, which coincides with $\mathcal{M}^\perp(H) = \mathcal{V}^\perp$. □

Proposition

Let $\mathcal{V}$ be a subspace and H be a projection onto $\mathcal{V}$. Then $Hy = y$ for all $y \in \mathcal{V}$.

Proposition

If P and Q are projection matrices onto a subspace $\mathcal{V}$, then $P = Q$.

Proposition

If $x_1, \ldots, x_p$ are linearly independent and are such that span$(x_1, \ldots, x_p) = \mathcal{V}$, then the projection onto $\mathcal{V}$ can be represented as

$$H = X(X^\top X)^{-1} X^\top$$

where X is a matrix with columns $x_1, \ldots, x_p$.

Proposition

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$ and H be a projection onto $\mathcal{V}$. Then

$$\|x - Hx\| \leq \|x - v\|, \qquad \forall v \in \mathcal{V}.$$

Proof

Let $H = U\Lambda U^\top$ be the spectral decomposition of H, $U = (u_1 \ \cdots \ u_n)$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Letting $p = \dim(\mathcal{V})$,

1. $\lambda_1 = \cdots = \lambda_p = 1$ and $\lambda_{p+1} = \cdots = \lambda_n = 0$,
2. $u_1, \ldots, u_n$ is an orthonormal basis of $\mathbb{R}^n$,
3. $u_1, \ldots, u_p$ is an an orthonormal basis of $\mathcal{V}$.

$$
\begin{aligned}
\|x - Hx\|^2 &= \sum_{i=1}^{n} (x^\top u_i - (Hx)^\top u_i)^2 && \text{[orthonormal basis]} \\
&= \sum_{i=1}^{n} (x^\top u_i - x^\top H u_i)^2 && \text{[H is symmetric]} \\
&= \sum_{i=1}^{n} (x^\top u_i - \lambda_i x^\top u_i)^2 && \text{[u's are eigenvectors of H]} \\
&= 0 + \sum_{i=p+1}^{n} (x^\top u_i)^2 && \text{[eigenvalues 0 or 1]} \\
&\leq \sum_{i=1}^{p} (x^\top u_i - v^\top u_i)^2 + \sum_{i=p+1}^{n} (x^\top u_i)^2 \quad \forall v \in \mathcal{V} = \text{span}\{u_1, ..., u_p\} \\
&= \sum_{i=1}^{p} (x^\top u_i - v^\top u_i)^2 + \sum_{i=p+1}^{n} (x^\top u_i - v^\top u_i)^2 \quad [v \in \text{span}\{u_1, ..., u_p\}] \\
&= \|x - v\|^2.
\end{aligned}
$$

□

Proposition

Let $\mathcal{V}_1 \subseteq \mathcal{V} \subseteq \mathbb{R}^n$ be two nested linear subspaces. If H_1 is the projection onto $\mathcal{V}_1$ and H is the projection onto $\mathcal{V}$, then

$$HH_1 = H_1 = H_1 H\,.$$

Proof.

First we show that $HH_1 = H_1$, and then that $H_1 H = HH_1$. For all $y \in \mathbb{R}^n$ we have $H_1 y \in \mathcal{V}_1$. But then $H_1 y \in \mathcal{V}$, since $\mathcal{V}_1 \subseteq \mathcal{V}$. Therefore $HH_1 y = H_1 y$. We have shown that $(HH_1 - H_1)y = 0$ for all $y \in \mathbb{R}^n$, so that $HH_1 - H_1 = 0$, as its kernel is all $\mathbb{R}^n$. Hence $HH_1 = H_1$.

(Or, take n linearly independent vectors $y_1, \ldots, y_n \in \mathbb{R}^n$, and use them as columns of the $n \times n$ matrix Y. Now Y is invertible, and $(HH_1 - H_1)Y = 0$, so $HH_1 - H_1 = 0$, giving $HH_1 = H_1$.)

To prove that $H_1 H = HH_1$, note that symmetry of projection matrices and the first part of the proof give

$$H_1 H = H_1^\top H^\top = (HH_1)^\top = (H_1)^\top = H_1 = HH_1.$$

□

Definition (Non-Negative Matrix – Quadratic Form Definition)

A $p \times p$ real symmetric matrix Ω is called non-negative definite (written $\Omega \succeq 0$) if and only if $x^\top \Omega x \geq 0$ for all $x \in \mathbb{R}^p$. If $x^\top \Omega x > 0$ for all $x \in \mathbb{R}^p \setminus \{0\}$, then we call Ω positive definite (written $\Omega \succ 0$).

An equivalent definition is:

Definition (Non-Negative Matrix – Spectral Definition)

A $p \times p$ real symmetric matrix Ω is called non-negative definite (written $\Omega \succeq 0$) if and only the eigenvalues of Ω are non-negative. If the eigenvalues of Ω are strictly positive, then Ω is called positive definite (written $\Omega \succ 0$).

Lemma (Exercise)

Prove that the two definitions are equivalent.

Definition (Covariance Matrix)

Let $Y = (Y_1, \ldots, Y_n)^\top$ be a random $n \times 1$ vector such that $\mathbb{E}\|Y\|^2 < \infty$. The covariance matrix of Y, say Ω, is the $n \times n$ symmetric matrix with entries

$$\Omega_{ij} = \text{cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])], \quad 1 \leq i \leq j \leq n.$$

That is, the covariance matrix encodes the variances of the coordinates of Y (on the diagonal) and the covariances between the coordinates of Y (off the diagonal). If we write

$$\mu = \mathbb{E}[Y] = (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n])^\top$$

for the mean vector of Y, then the covariance matrix of Y can be written as

$$\mathbb{E}[(Y - \mu)(Y - \mu)^\top] = \mathbb{E}[YY^\top] - \mu\mu^\top.$$

Whenever Y is a random vector, we will write $\text{cov}(Y)$ or $\text{var}(Y)$ for the covariance matrix of Y.

Lemma

Let Y be a random $d \times 1$ vector such that $\mathbb{E}\|Y\|^2 < \infty$. Let μ be the mean vector and Ω be the covariance matrix of Y. If A is a $p \times d$ real matrix, the mean vector and covariance matrix of AY are $A\mu$ and $A\Omega A^\top$, respectively.

Proof.

Exercise. □

Corollary (Covariance of Projections)

Let Y be a random $d \times 1$ vector such that $\mathbb{E}\|Y\|^2 < \infty$. Let $\beta, \gamma \in \mathbb{R}^d$ be fixed vectors. If Ω denotes the covariance matrix of Y,

- *the variance of $\beta^\top Y$ is $\beta^\top \Omega \beta$;*
- *the covariance of $\beta^\top Y$ with $\gamma^\top Y$ is $\gamma^\top \Omega \beta$.*

Proposition (Non-Negative and Covariance Matrices)

Let Ω be a real symmetric matrix. Then Ω is non-negative definite if and only if Ω is the covariance matrix of some random variable Y.

Proof.

Exercise. $\square$

- Let Y be a random vector in $\mathbb{R}^d$ with covariance matrix Ω.
- Find direction $v_1 \in \mathbb{S}^{d-1}$ such that the projection of Y onto v_1 has maximal variance.
- For $j = 2, 3, \ldots, d$, find direction $v_j \perp v_{j-1}$ such that projection of Y onto v_j has maximal variance.

Solution: maximise $\text{Var}(v_1^\top Y) = v_1^\top \Omega v_1$ over $\|v_1\| = 1$

$$v_1^\top \Omega v_1 = v_1^\top U \Lambda U^\top v_1 = \|\Lambda^{1/2} U^\top v_1\|^2 = \sum_{i=1}^{d} \lambda_i (u_i^\top v_1)^2 \qquad \text{[change of basis]}$$

Now $\sum_{i=1}^{d} (u_i^\top v_1)^2 = \|v_1\|^2 = 1$ so we have a convex combination of the $\{\lambda_j\}_{j=1}^{d}$,

$$\sum_{i=1}^{d} p_i \lambda_i, \qquad \sum_i p_i = 1, \quad p_i \geq 0, \quad i = 1, \ldots, d.$$

But $\lambda_1 \geq \lambda_i \geq 0$ so clearly this sum is maximised when $p_1 = 1$ and $p_j = 0$ $\forall j \neq 1$, i.e. $v_1 = \pm u_1$.

Iteratively, $v_j = \pm u_j$, i.e. principal components are eigenvectors of Ω.

Theorem (Optimal Linear Dimension Reduction Theorem)

Let Y be a mean-zero random variable in $\mathbb{R}^n$ with $n \times n$ covariance Ω. Let H be the projection matrix onto the span of the first k eigenvectors of Ω. Then

$$\mathbb{E}\|Y - HY\|^2 \leq \mathbb{E}\|Y - QY\|^2$$

for any $n \times n$ projection operator Q or rank at most k.

Intuitively: if you want to approximate a mean-zero random variable taking values $\mathbb{R}^n$ by a random variable that ranges over a subspace of dimension at most $k \leq n$, the optimal choice is the projection of the random variable onto the space spanned by its first k principal components (eigenvectors of the covariance). "Optimal" is with respect to the mean squared error.

For the proof, use lemma below (follows immediately from spectral decomposition)

Lemma

Q is a rank k projection matrix if and only if there exist orthonormal vectors $\{v_j\}_{j=1}^k$ such that $Q = \sum_{j=1}^k v_j v_j^\top$.

Optimal Linear Dimension Reduction.

Write $Q = \sum_{j=1}^{k} v_j v_j^\top$ for some orthonormal $\{v_j\}_{j=1}^{k}$. Then,

$$\begin{aligned}
\mathbb{E}\|Y - QY\|^2 &= \mathbb{E}\left[Y^\top (I-Q)^\top (I-Q) Y\right] = \mathbb{E}\left[\text{tr}\{(I-Q) Y Y^\top (I-Q)^\top\}\right] \\
&= \text{tr}\{(I-Q)\mathbb{E}\left[YY^\top\right](I-Q)^\top\} = \text{tr}\{(I-Q)^\top (I-Q)\Omega\} \\
&= \text{tr}\{(I-Q)\Omega\} = \text{tr}\{\Omega\} - \text{tr}\{Q\Omega\} = \sum_{i=1}^{n}\lambda_i - \text{tr}\left\{\sum_{j=1}^{k} v_j v_j^\top \Omega\right\} \\
&= \sum_{i=1}^{n}\lambda_i - \sum_{j=1}^{k}\text{tr}\left\{v_j v_j^\top \Omega\right\} = \sum_{i=1}^{n}\lambda_i - \sum_{j=1}^{k} v_j^\top \Omega v_j \\
&= \sum_{i=1}^{n}\lambda_i - \sum_{j=1}^{k}\text{Var}[v_j^\top Y]
\end{aligned}$$

If we can minimise this expression over all $\{v_j\}_{j=1}^{k}$ with $v_j^\top v_{j'} = \mathbf{1}\{j = j'\}$, then we're done. By PCA, this is done by choosing the top k eigenvectors of Ω. □

Corollary

Let $\{x_1, ..., x_p\} \subset \mathbb{R}^n$ be such that $x_1 + \ldots + x_p = 0$, and let X be the $n \times p$ matrix with columns $\{x_j\}_{j=1}^p$. The best approximating k-hyperplane to the points $\{x_1, ..., x_p\}$ is given by the span of the k leading eigenvectors of the matrix $XX^\top$, i.e. if H is the projection onto this span, it holds that

$$\sum_{j=1}^{p} \|x_j - Hx_j\|^2 \leq \sum_{j=1}^{p} \|x_j - Qx_j\|^2$$

for any $n \times n$ projection operator Q or rank at most k.

Proof.

Define a discrete random vector Y by $\mathbb{P}[Y = x_j] = 1/p$, $j \in \{1, ..., p\}$ and observe that $\mathbb{E}[h(Y)] = p^{-1} \sum_{j=1}^{p} h(x_j)$, for any vector-valued (or matrix-valued) deterministic map h. Now use the optimal linear dimension reduction theorem. □

Definition (Multivariate Gaussian Distribution)

A random vector Y in $\mathbb{R}^d$ has the multivariate normal distribution if and only if $\beta^\top Y$ has the univariate normal distribution, $\forall \beta \in \mathbb{R}^d$.

Observation: From the definition if follows that Y must have some well-defined mean vector μ and some well defined covariance matrix Ω.

To see this note that since $\mathbb{E}\{(\beta^\top Y)^2\} < \infty$ for all β, then we can successively pick β to be equal to each canonical basis vector and conclude that each coordinate has finite variance and thus $\mathbb{E}\| Y \|^2 < \infty$.

So all the means, variances and covariances of its coordinates are well defined.

Then, the mean vector (say) μ and covariance matrix (say) Ω can be (uniquely) determined entrywise by equating

$$\mu_i = \mathbb{E}[e_i^\top Y] \qquad \& \qquad \Omega_{ij} = \text{cov}\{e_i^\top Y,\, e_j^\top Y\}.$$

where e_j is the jth canonical basis vector

$$e_j = (0\ ,\ 0\ , \ldots, \underbrace{1}_{j^{th}\ \text{position}}, \ldots,\ 0\ ,\ 0)^\top$$

How can we use this definition to determine basic properties?

The *moment generating function* (MGF) of a random vector W in $\mathbb{R}^d$ is defined as

$$M_W(\theta) = \mathbb{E}[e^{\theta^\top W}], \qquad \theta \in \mathbb{R}^d,$$

provided the expectation exists. When the MGF exists *it characterises the distribution of the random vector. Furthermore, two random vectors are independent if and only if their joint MGF is the product of their marginal MGF's*, i.e.

$$\boxed{X_{n\times 1} \text{ independent of } Y_{m\times 1}}$$

$$\Longleftrightarrow$$

$$\boxed{\mathbb{E}[e^{\beta^\top X+\gamma^\top Y}] = \mathbb{E}[e^{\beta^\top X}] \times \mathbb{E}[e^{\gamma^\top Y}], \qquad \forall\, \beta \in \mathbb{R}^n \,\&\, \gamma \in \mathbb{R}^m}$$

Gaussian vector basic factsheet:

1. Moment generating function of $Y \sim \mathcal{N}(\mu, \Omega)$:

$$\boxed{M_Y(u) = \exp\left(u^\top \mu + \frac{1}{2} u^\top \Omega u\right).}$$

2. $Y \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$ and given $B_{n\times p}$ and $\theta_{n\times 1}$, then
$$\theta + BY \sim \mathcal{N}(\theta + B\mu, B\Omega B^\top).$$
3. $\mathcal{N}(\mu, \Omega)$ density, assuming Ω nonsingular:

$$\boxed{f_Y(y) = \frac{1}{(2\pi)^{p/2}\,|\Omega|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu)^\top \Omega^{-1}(y-\mu)\right\}.}$$

4. Constant density isosurfaces are ellipsoidal
5. Marginals of Gaussian are Gaussian (converse NOT true).
6. Ω diagonal $\Leftrightarrow$ independent coordinates Y_j.
7. If $Y \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$,
$$AY \text{ independent of } BY \iff A\Omega B^\top = 0.$$

Proposition (Property 1: Moment Generating Function)

The moment generating function of $Y \sim \mathcal{N}(\mu, \Omega)$ *is*

$$M_Y(u) = \exp\left(u^\top \mu + \tfrac{1}{2} u^\top \Omega u\right)$$

Proof.

Let $v \in \mathbb{R}^d$ be arbitrary. Then $v^\top Y$ is scalar Gaussian with mean $v^\top \mu$ and variance $v^\top \Omega v$. Hence it has moment generating function:

$$M_{v^\top Y}(t) = \mathbb{E}\left(e^{tv^\top Y}\right) = \exp\left\{t(v^\top \mu) + \frac{t^2}{2}(v^\top \Omega v)\right\}.$$

Now take $t = 1$ and observe that

$$M_{v^\top Y}(1) = \mathbb{E}\left(e^{v^\top Y}\right) = M_Y(v).$$

Combining the two, we conclude that

$$M_Y(v) = \exp\left(v^\top \mu + \frac{1}{2} v^\top \Omega v\right), \quad v \in \mathbb{R}^d.$$

□

Proposition (Property 2: Affine Transformation)

For $Y \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$ and given $B_{n\times p}$ and $\theta_{n\times 1}$, we have

$$\theta + BY \sim \mathcal{N}(\theta + B\mu, B\Omega B^\top)$$

Proof.

$$\begin{aligned}
M_{\theta+BY}(u) &= \mathbb{E}\left[\exp\{u^\top(\theta + BY)\}\right] = \exp\left\{u^\top\theta\right\}\mathbb{E}\left[\exp\{(B^\top u)^\top Y\}\right] \\
&= \exp\left\{u^\top\theta\right\} M_Y(B^\top u) \\
&= \exp\left\{u^\top\theta\right\} \exp\left\{(B^\top u)^\top\mu + \frac{1}{2}u^\top B\Omega B^\top u\right\} \\
&= \exp\left\{u^\top\theta + u^\top(B\mu) + \frac{1}{2}u^\top B\Omega B^\top u\right\} \\
&= \exp\left\{u^\top(\theta + B\mu) + \frac{1}{2}u^\top B\Omega B^\top u\right\}
\end{aligned}$$

And this last expression is the MGF of a $\mathcal{N}(\theta + B\mu, B\Omega B^\top)$ distribution. □

Proposition (Property 3: Density Function)

Let $\Omega_{p\times p}$ be nonsingular. The density of $\mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$ is

$$f_Y(y) = \frac{1}{(2\pi)^{p/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^\top \Omega^{-1}(y-\mu)\right\}$$

Proof.

We will first find the $\mathcal{N}(0, I)$, and then get the general one by change of variables (leveraging the last result).

To this aim, let $Z = (Z_1, \ldots, Z_p)^\top$ be a vector of iid $\mathcal{N}(0,1)$ random variables. Then, because of independence,

(a) the density of Z is

$$f_Z(z) = \prod_{i=1}^p f_{Z_i}(z_i) = \prod_{i=1}^p \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}z^\top z\right).$$

(b) The MGF of Z is

$$M_Z(u) = \mathbb{E}\left\{\exp\left(\sum_{i=1}^p u_i Z_i\right)\right\} = \prod_{i=1}^p \mathbb{E}\{\exp(u_i Z_i)\} = \exp(u^\top u/2),$$

which is the MGF of a p-variate $\mathcal{N}(0, I)$ distribution.

proof continued

$\overset{(a)+(b)}{\Longrightarrow}$ the $\mathcal{N}(0, I)$ density is $f_Z(z) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} z^\top z\right)$.

By the spectral theorem, Ω admits a non-negative definite square root, $\Omega^{1/2}$. Furthermore, since Ω is non-singular, so is $\Omega^{1/2}$.

Now observe that from our Property 2, we have $Y \overset{d}{=} \Omega^{1/2} Z + \mu \sim \mathcal{N}(\mu, \Omega)$.

By the change of variables formula,

$$\begin{aligned} f_Y(y) &= f_{\Omega^{1/2} Z + \mu}(y) \\ &= |\Omega^{-1/2}| f_Z\{\Omega^{-1/2}(y - \mu)\} \\ &= \frac{1}{(2\pi)^{p/2} |\Omega|^{1/2}} \exp\left\{-\frac{1}{2}(y - \mu)^\top \Omega^{-1}(y - \mu)\right\}. \end{aligned}$$

[Recall that to obtain the density of $W = g(X)$ at w, we need to evaluate f_X at $g^{-1}(w)$ but also multiply by the Jacobian determinant of g^{-1} at w.]

□

Proposition (Property 4: Isosurfaces)

The isosurfaces of a $\mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$ are $(p-1)$-dimensional ellipsoids centred at μ, with principal axes given by the eigenvectors of Ω and with anisotropies given by the ratios of the square roots of the corresponding eigenvalues of Ω .

Proof.

Exercise: Use Property 3, and the spectral theorem. □

Proposition (Property 5: Coordinate Distributions)

Let $Y = (Y_1, \ldots, Y_p)^\top \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$. Then $Y_j \sim \mathcal{N}(\mu_j, \Omega_{jj})$.

Proof.

Observe that $Y_j = (0\ ,\ 0\ , \ldots,\ \underbrace{1}_{j_{th}\text{ position}}\ , \ldots,\ 0\ ,\ 0)\, Y$ and use Property 2. □

Proposition (Property 6: Diagonal $\Omega \iff$ Independence)

Let $Y = (Y_1, \ldots, Y_p)^\top \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$. Then the Y_i are mutually independent if and only if Ω is diagonal.

Proof.

Suppose that the Y_j are independent. Property 5 yields $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for some $\sigma_j > 0$. Thus the density of Y is

$$f_Y(y) = \prod_{j=1}^{p} f_{Y_j}(y_j) = \prod_{i=1}^{p} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(y_j-\mu_j)^2}{\sigma_j^2}\right\}$$

$$= \frac{1}{(2\pi)^{p/2}\,|\mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu)^\top \mathrm{diag}(\sigma_1^{-2},\ldots,\sigma_p^{-2})(y-\mu)\right\}.$$

Hence $Y \sim \mathcal{N}\{\mu, \mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)\}$, i.e. the covariance Ω is diagonal.

Conversely, assume Ω is diagonal, say $\Omega = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)$. Then we can reverse the steps of the first part to see that the joint density $f_Y(y)$ can be written as a product of the marginal densities $f_{Y_j}(y_j)$, thus proving independence.

□

Proposition (Property 7: AY, BY indep $\iff$ $A\Omega B^\top = 0$)

If $Y \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$, and $A_{m\times p}$, $B_{d\times p}$ be real matrices. Then,

$$AY \text{ independent of } BY \iff A\Omega B^\top = 0.$$

Proof

It suffices to prove the result assuming $\mu = 0$ (and it simplifies the algebra).
First assume $A\Omega B^\top = 0$. Let $W_{(m+d)\times 1} = \binom{AY}{BY}$ and $\theta_{(m+d)\times 1} = \binom{u_{m\times 1}}{v_{d\times 1}}$.

$$\begin{aligned}
M_W(\theta) &= \mathbb{E}[\exp\{W^\top\theta\}] = \mathbb{E}\left[\exp\left\{Y^\top A^\top u + Y^\top B^\top v\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{Y^\top (A^\top u + B^\top v)\right\}\right] = M_Y(A^\top u + B^\top v) \\
&= \exp\left\{\frac{1}{2}(A^\top u + B^\top v)^\top \Omega (A^\top u + B^\top v)\right\} \\
&= \exp\left\{\frac{1}{2}\left(u^\top A\Omega A^\top u + v^\top B\Omega B^\top v + u^\top \underbrace{A\Omega B^\top}_{=0} v + v^\top \underbrace{B\Omega A^\top}_{=0} u\right)\right\} \\
&= M_{AY}(u) M_{BY}(v),
\end{aligned}$$

i.e., the joint MGF is the product of the marginal MGFs, proving independence.

For the converse, assume that AY and BY are independent. Then, $\forall u, v$,

$$M_W(\theta) = M_{AY}(u) M_{BY}(v), \quad \forall u, v,$$

$$\implies \exp\left\{\frac{1}{2}\left(u^\top A\Omega A^\top u + v^\top B\Omega B^\top v + u^\top A\Omega B^\top v + v^\top B\Omega A^\top u\right)\right\}$$

$$= \exp\left\{\frac{1}{2} u^\top A\Omega A^\top u\right\} \exp\left\{\frac{1}{2} v^\top B\Omega B^\top v\right\}$$

$$\implies \exp\left\{\frac{1}{2} \times 2u^\top A\Omega B^\top v\right\} = 1$$

$$\implies u^\top A\Omega B^\top v = 0, \qquad \forall\ u \in \mathbb{R}^d, v \in \mathbb{R}^m,$$

$\implies$ the orthocomplement[a] of the column space of $A\Omega B^\top$ is the whole of $\mathbb{R}^m$.

$\implies$ the column space of $A\Omega B^\top$ is the trivial subspace $\{0\}$.

$$\implies A\Omega B^\top = 0.$$

□

[a] recall that for $Q_{m\times d}$ we have $\mathcal{M}^\perp(Q) = \{y \in \mathbb{R}^m : y^\top Q x = 0,\ \forall x \in \mathbb{R}^d\}$

Definition (χ^2 distribution)

Let $Z \sim \mathcal{N}(0, I_{p\times p})$. Then $\|Z\|^2 = \sum_{j=1}^{p} Z_j^2$ is said to have the chi-square (χ^2) distribution with p degrees of freedom; we write $\|Z\|^2 \sim \chi^2_p$.

[Thus, χ^2_p is the distribution of the sum of squares of p real independent standard Gaussian random variates.]

Definition (F distribution)

Let $V \sim \chi^2_p$ and $W \sim \chi^2_q$ be independent random variables. Then $(V/p)/(W/q)$ is said to have the F distribution with p and q degrees of freedom; we write $(V/p)/(W/q) \sim F_{p,q}$.

Proposition (Gaussian Quadratic Forms)

1. *If* $Z \sim \mathcal{N}(0_{p\times 1}, I_{p\times p})$ *and* H *is a projection of rank* $r \leq p$,

$$Z^\top H Z \sim \chi^2_r.$$

2. $Y \sim \mathcal{N}(\mu_{p\times 1}, \Omega_{p\times p})$ *with* Ω *nonsingular* $\Longrightarrow$

$$(Y-\mu)^\top \Omega^{-1}(Y-\mu) \sim \chi^2_p.$$

Exercise: Prove these results.

What if the random vector is not Gaussian? Here's a CLT[2] that helps:

Theorem (Hajék-Šidák Weighted Sum CLT)

Let $\{X_n\}$ be an i.i.d sequence of real random variables, with common mean 0 and variance 1. Let $\{\gamma_n\}$ be a sequence of real constants. Then,

$$\sup_{1\leq j\leq n} \frac{\gamma_j^2}{\sum_{i=1}^n \gamma_i^2} \overset{n\to\infty}{\longrightarrow} 0 \Longrightarrow \frac{1}{\sqrt{\sum_{i=1}^n \gamma_i^2}} \sum_{i=1}^n \gamma_i X_i \overset{d}{\to} N(0,1).$$

- Supremum condition amounts to saying that, in the limit, any single component contributes a negligible proportion of the total variance.
- Coefficient sequence $\{\gamma_n\}$ might very well diverge, without contradicting the negligibility condition (e.g. $\gamma_k = \sqrt{k}$)

[2] Consequence of Lyapunov's CLT, see e.g. Sen & Singer, "Large Sample Methods in Statistics", Chapman & Hall, pp. 108-119.

Linear Models: Likelihood and Geometry

General formulation:

$$Y_i | x_i \overset{ind}{\sim} \text{Distribution}\{g(x_i)\}, \quad i = 1, \ldots, n.$$

Simple Normal Linear Regression:

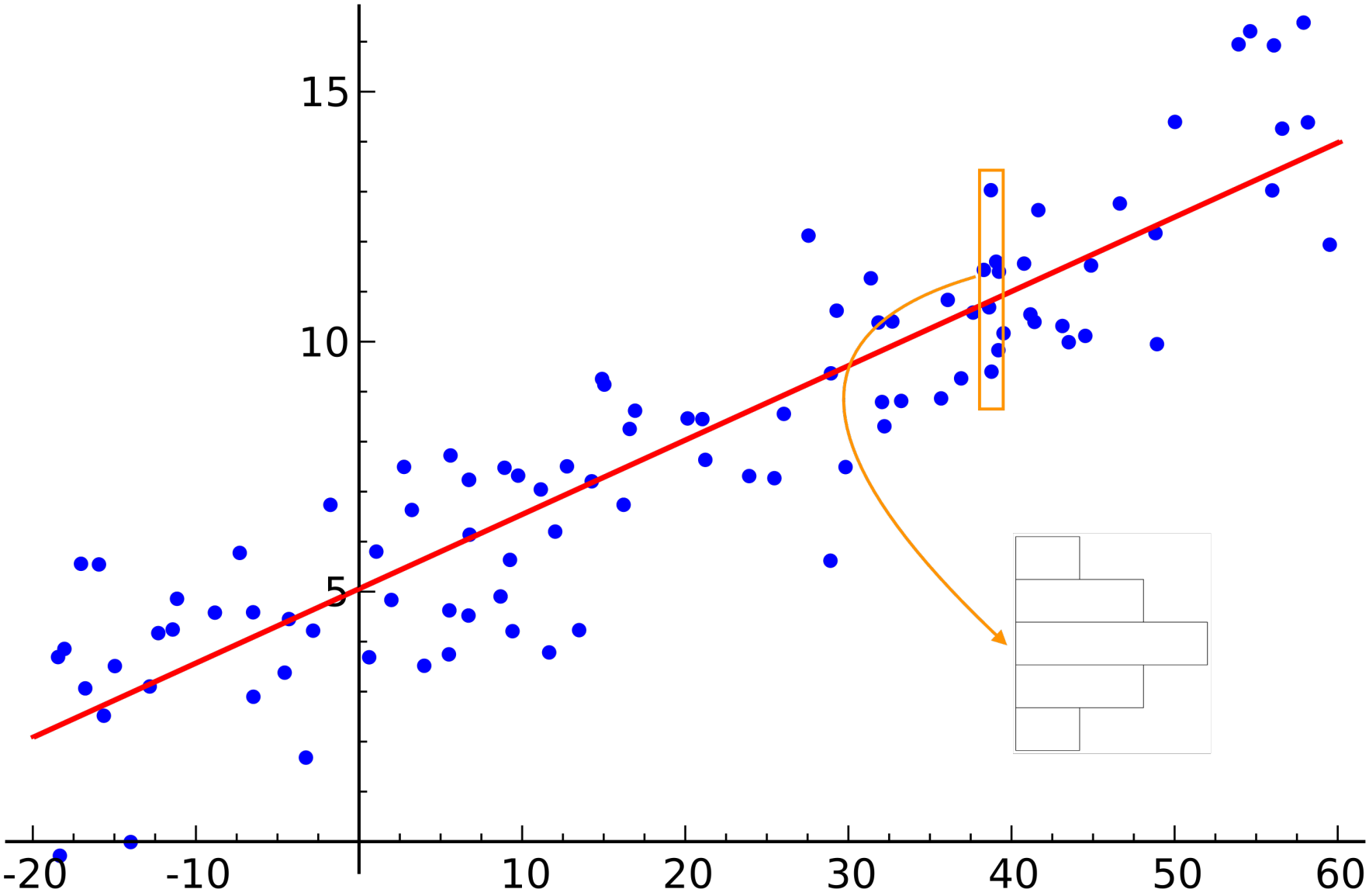
$$\begin{cases} \text{Distribution} = \mathcal{N}\{g(x), \sigma^2\} \\ g(x) = \beta_0 + \beta_1 x \end{cases}$$

Resulting Model:

$$Y_i \overset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

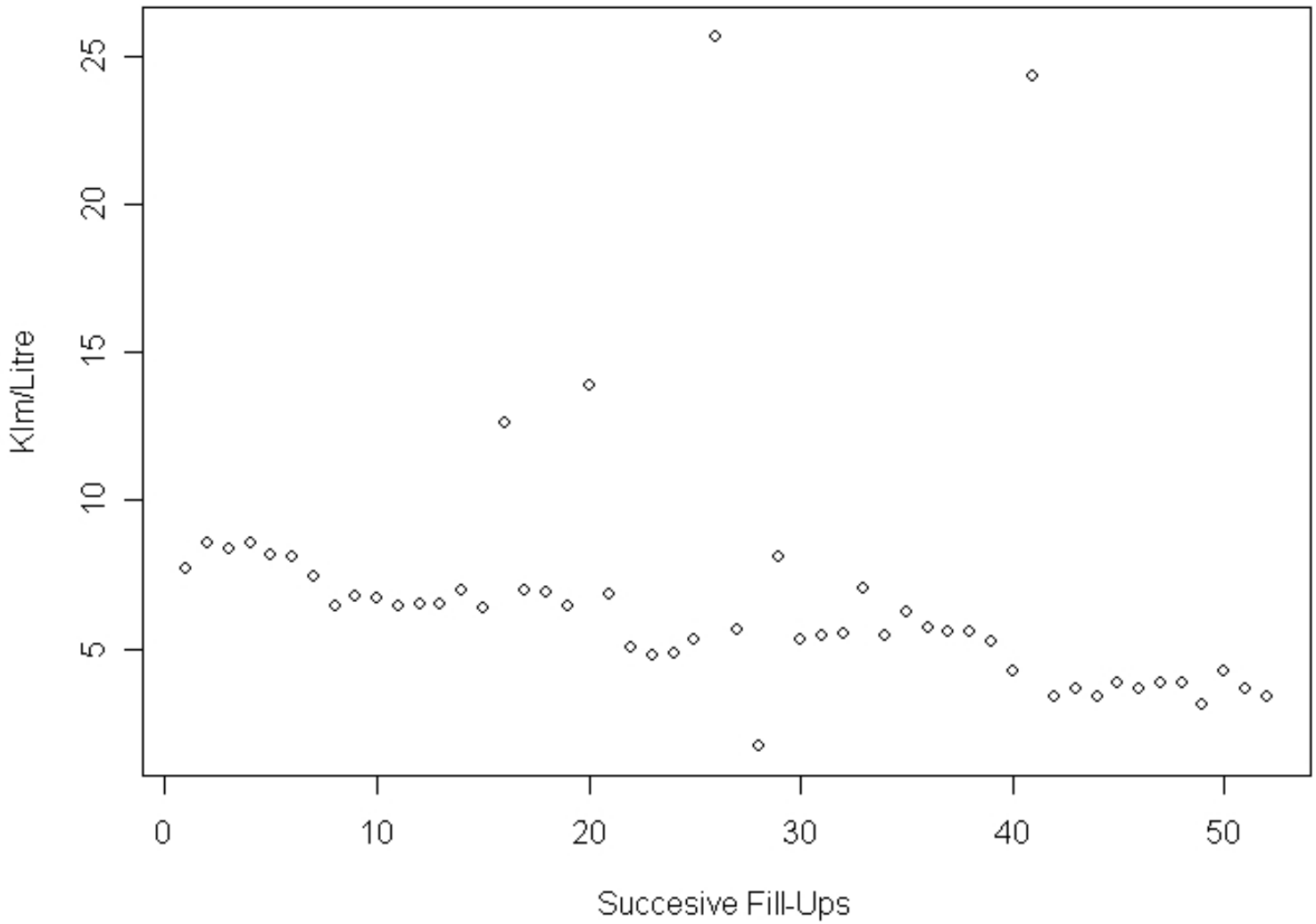
$$\Updownarrow$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$$


15
10
5
-20
-10
10
20
30
40
50
60

Fillup	Km/L
1	7.72
2	8.54
3	8.35
4	8.55
5	8.16
6	8.12
7	7.46
8	6.43
9	6.74
10	6.72

Example: Professor's Van



Jargon: Y is *response variable* and x is *explanatory variable* (or *covariate*)
Linearity: Linearity is in the *parameters*, not the *explanatory variable*.
Example: Flexibility in what we define as explanatory:

$$Y_j = \beta_0 + \beta_1 \underbrace{\sin(x_j)}_{x_j^*} + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \text{Normal}(0, \sigma^2).$$

Example: Sometimes a transformation may be required:

$$Y_j = \beta_0 \, e^{\beta_1 x_j} \eta_j, \quad \eta_j \overset{iid}{\sim} \text{Lognormal}$$

$$\log(\cdot) \downarrow \qquad \uparrow \exp(\cdot)$$

$$\log Y_j = \log \beta_0 + \beta_1 x_j + \log \eta_j, \quad \log \eta_j \overset{iid}{\sim} \text{Normal}$$

Data Structure:
For $i = 1, \ldots, n$, pairs

$$(x_i, y_i) \longrightarrow \begin{cases} x_i \text{ fixed values of } x \\ y_i \text{ treated as a realisation of } Y_i \text{ at } x_i \end{cases}$$

Instead of $x_i \in \mathbb{R}$ could have $x_i^\top \in \mathbb{R}^q$):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2).$$

Letting $p = q + 1$, this can be summarised via matrix notation:

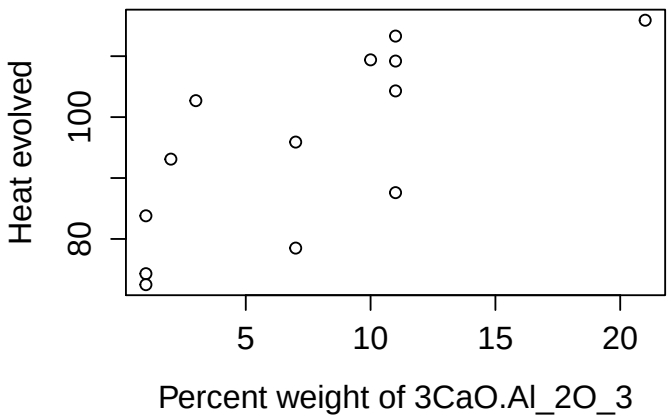
$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1q} \\ 1 & x_{21} & & x_{2q} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nq} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

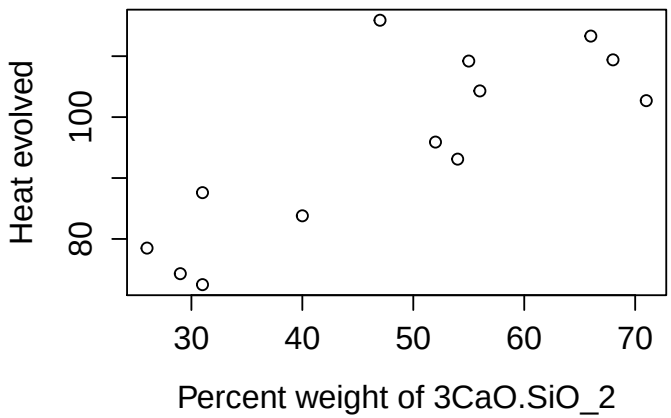
$$\implies \underset{n\times 1}{Y} = \underset{n\times p}{X} \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

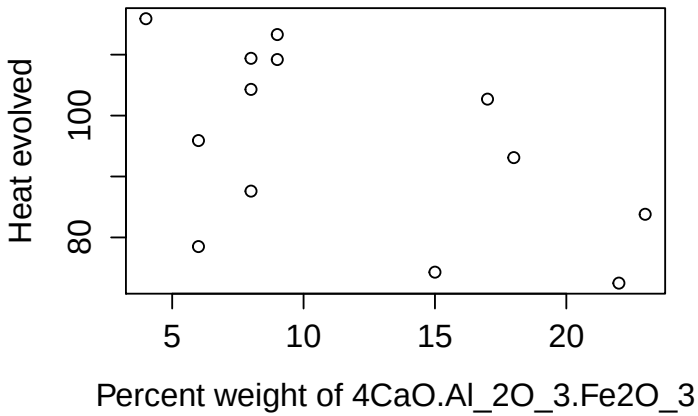
X is called the *design matrix*.

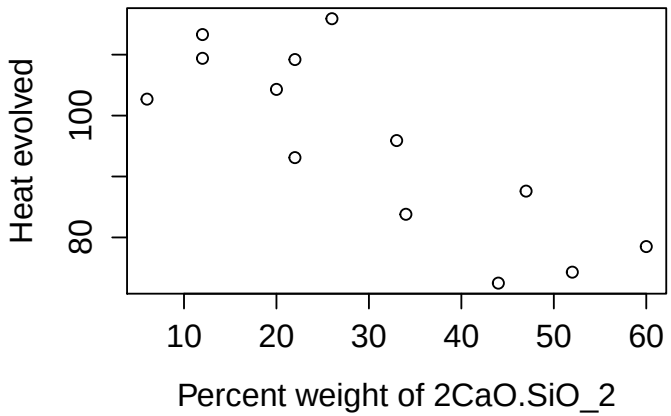
Case	$3\,CaO.Al_2O_3$	$3\,CaO.SiO_2$	$4\,Cao.Al_2O_3.Fe_2O_3$	$2\,CaO.SiO_2$	Heat
1	7.00	26.00	6.00	60.00	78.50
2	1.00	29.00	15.00	52.00	74.30
3	11.00	56.00	8.00	20.00	104.30
4	11.00	31.00	8.00	47.00	87.60
5	7.00	52.00	6.00	33.00	95.90
6	11.00	55.00	9.00	22.00	109.20
7	3.00	71.00	17.00	6.00	102.70
8	1.00	31.00	22.00	44.00	72.50
9	2.00	54.00	18.00	22.00	93.10
10	21.00	47.00	4.00	26.00	115.90
11	1.00	40.00	23.00	34.00	83.80
12	11.00	66.00	9.00	12.00	113.30
13	10.00	68.00	8.00	12.00	109.40

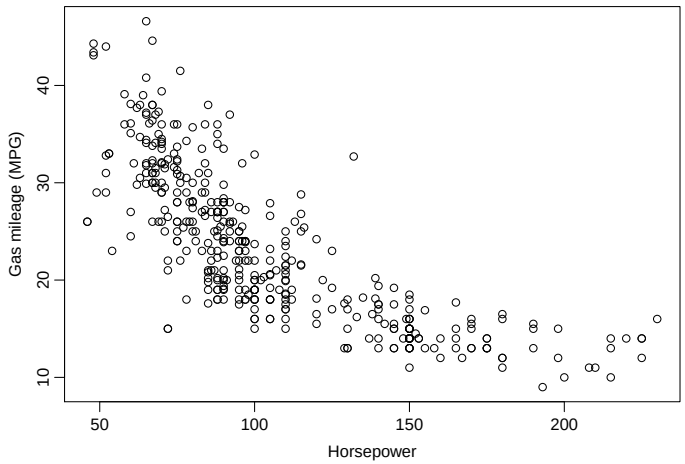
Cement Heat Evolution











Perhaps more fitting than

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

would be

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \varepsilon_j$$

Still a linear model but now with 2 covariates: x_j and $x_j^* = x_j^2$

- Normally would require a (hyper)plane to visualise dependence of mean on 2 or more covariates
- When additional covariates are variable transformation, can visualise mean dependence via a non-linear curve, even though model is linear

Model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\Updownarrow$$

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

Observe: $y = (y_1, \ldots, y_n)^\top$ for given fixed design matrix X, i.e.:

$$(y_1, x_{11}, \ldots, x_{1q}), \ldots, (y_i, x_{i1}, \ldots, x_{iq}), \ldots, (y_n, x_{n1}, \ldots, x_{nq})$$

Likelihood and Loglikelihood

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right\}$$

$$\ell(\beta, \sigma^2) = -\frac{1}{2}\left\{n \log 2\pi + n \log \sigma^2 + \frac{1}{\sigma^2}(y - X\beta)^\top (y - X\beta)\right\}$$

Whatever the value of σ, the log-likelihood is maximised when $(y - X\beta)^\top (y - X\beta)$ is minimised. Hence, the MLE of β is:

$$\hat{\beta} = \arg\max_{\beta} \left\{ -(y - X\beta)^\top (y - X\beta) \right\} = \arg\min_{\beta} (y - X\beta)^\top (y - X\beta)$$

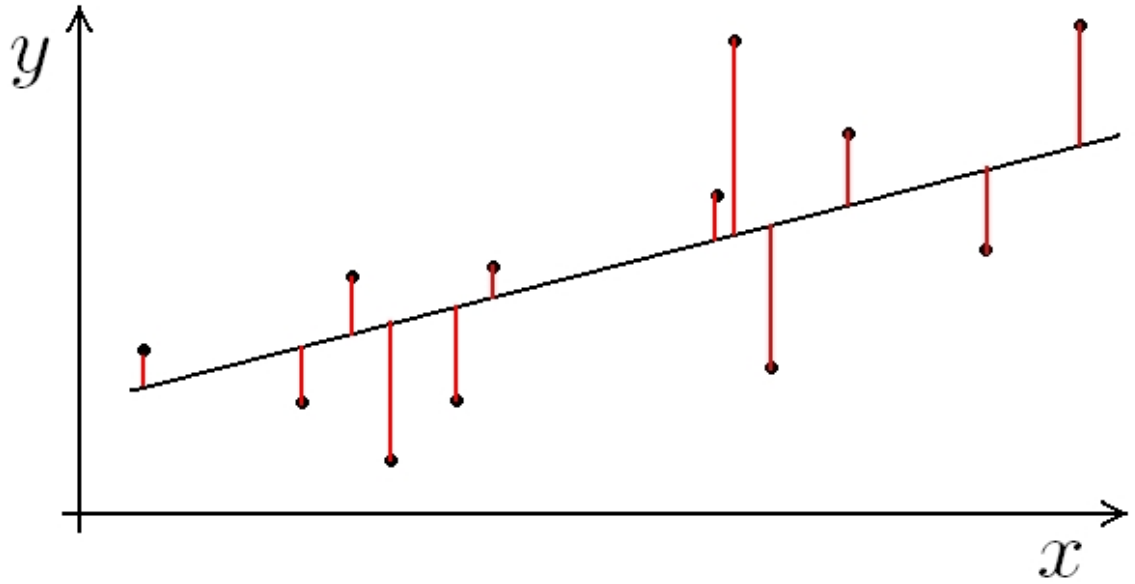
Obtain minimum by solving:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \beta} (y - X\beta)^\top (y - X\beta) \\
0 &= \frac{\partial (y - X\beta)}{\partial \beta} \, \frac{\partial (y - X\beta)^\top (y - X\beta)}{\partial (y - X\beta)} \quad \text{(chain rule)} \\
0 &= X^\top (y - X\beta) \quad \text{(normal equations)} \\
X^\top X\beta &= X^\top y \\
\hat{\beta} &= (X^\top X)^{-1} X^\top y \quad \text{(if } X \text{ has rank } p\text{)}
\end{aligned}$$

$\hat{\beta}$ is called the *least squares estimator* because it is a result of minimising

$$(y - X\beta)^\top (y - X\beta) = \underbrace{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_q x_{iq})^2}_{\text{sum of squares}}.$$

Thus we are trying to find the β that gives the hyperplane with minimum sum of squared vertical distances from our observations.



Residuals: $e = y - X\hat{\beta}$, so that $e = (e_1, \ldots, e_n)^\top$, with

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_q x_{iq}$$

"Regression Line" is such that $\sum e_i^2$ is minimised over all β.

Fitted Values: $\hat{y} = X\hat{\beta}^\top$, so that $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)^\top$, with

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{iq}$$

Since the MLE of β is $\hat{\beta} = (X^\top X)^{-1} X^\top y$ for all values of σ^2, we have

$$\begin{aligned}
\hat{\sigma}^2 &= \arg\max_{\sigma^2} \left\{ \max_{\beta} \ell(\beta, \sigma^2) \right\} \\
&= \arg\max_{\sigma^2} \ell(\hat{\beta}, \sigma^2) \\
&= \arg\max_{\sigma^2} \; -\frac{1}{2} \left\{ n \log \sigma^2 + \frac{1}{\sigma^2} (y - X\hat{\beta})^\top (y - X\hat{\beta}) \right\}.
\end{aligned}$$

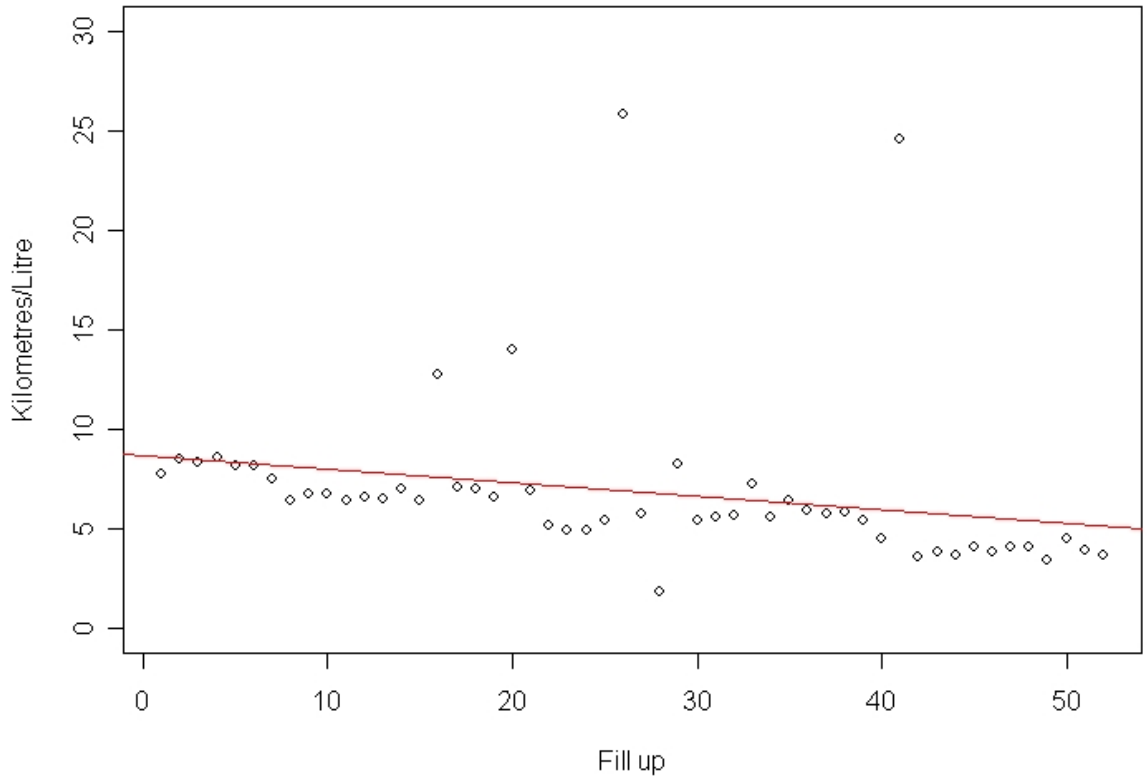
Differentiating and setting equal to zero yields

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^\top (y - X\hat{\beta}).$$

Next week we will see that a better (unbiased) estimator is

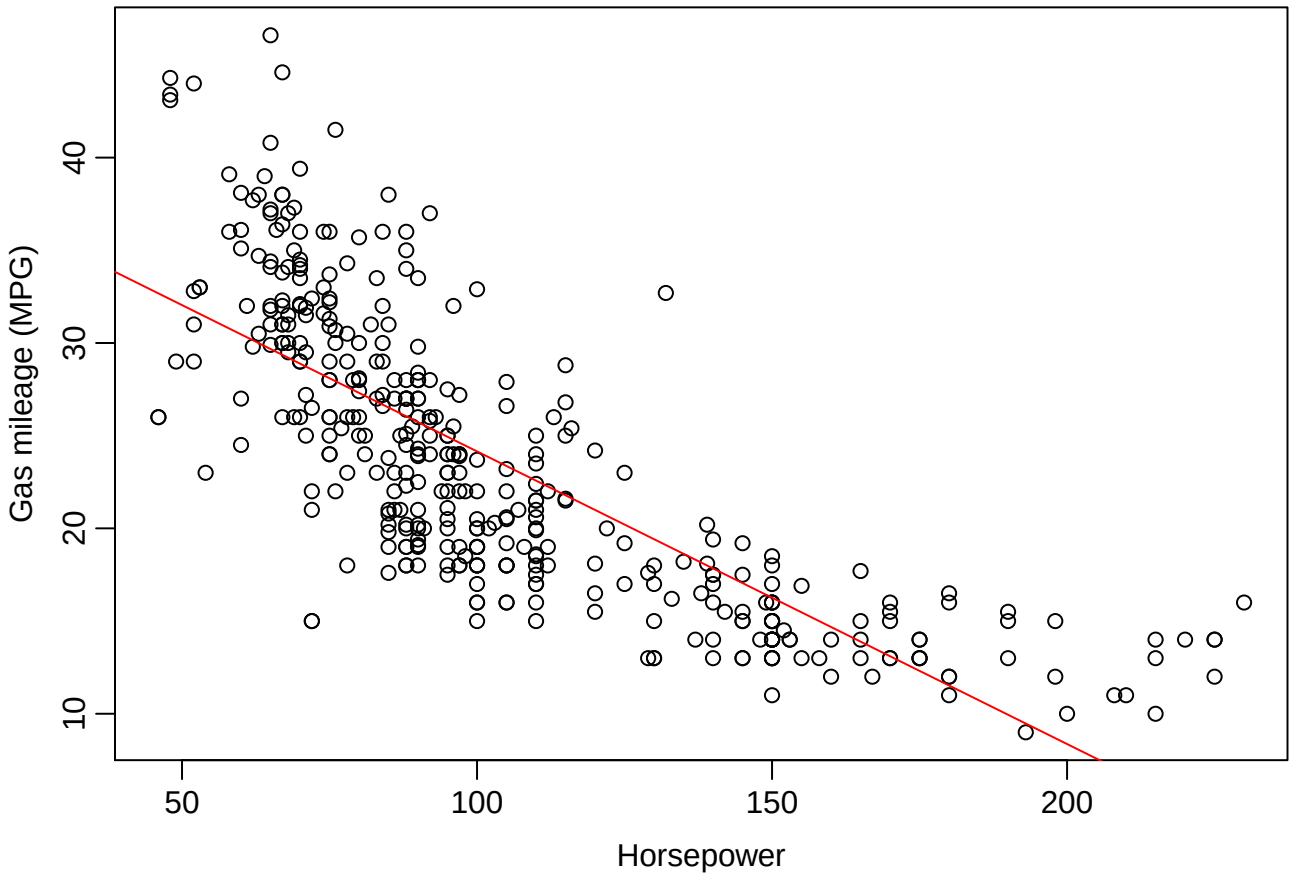
$$S^2 = \frac{1}{n-p} (y - X\hat{\beta})^\top (y - X\hat{\beta}).$$

Example: Professor's Van



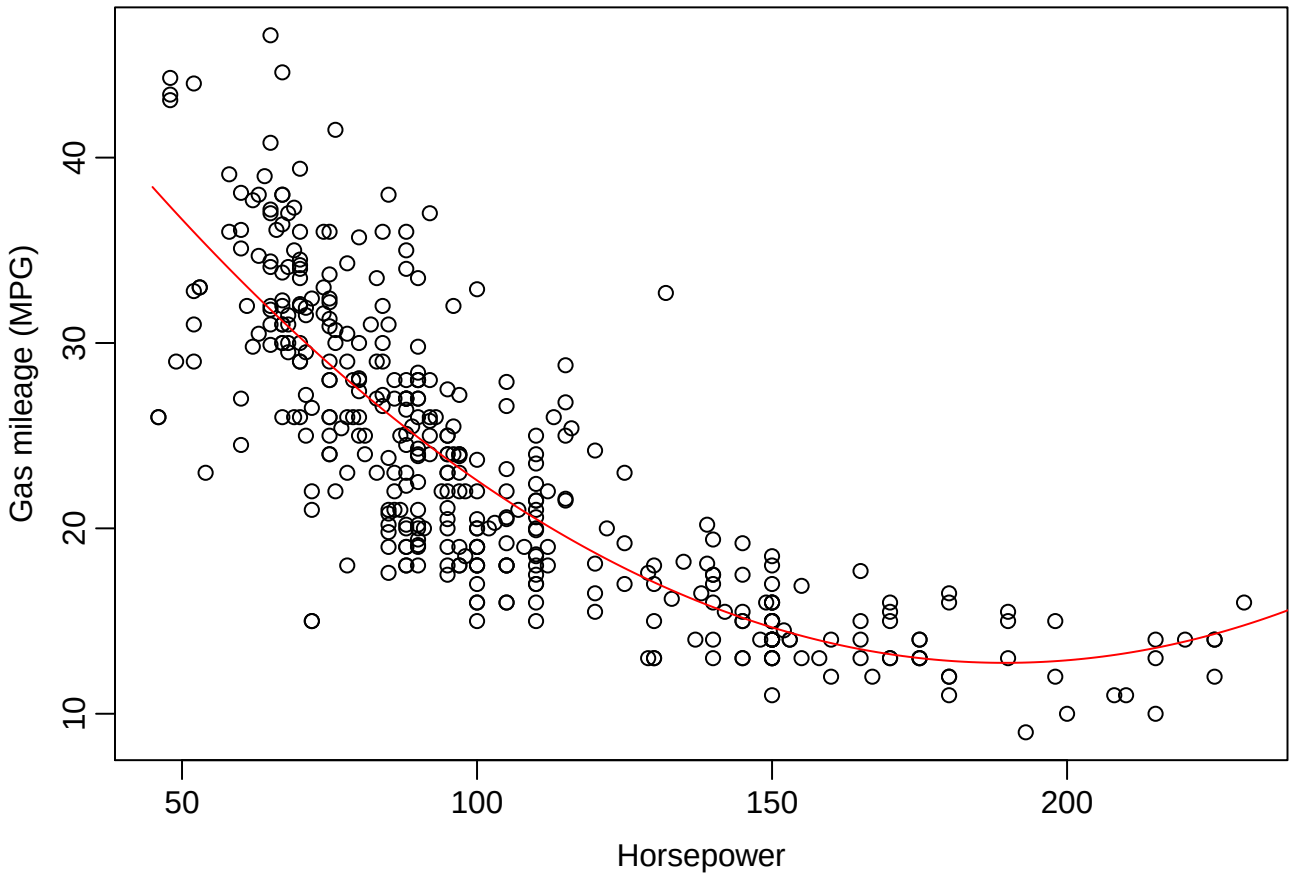
$$\hat{\beta}_0 = 8.6 \quad \hat{\beta}_1 = -0.068 \quad S^2 = 17.4$$

Model with linear term only



Parameter estimates: $\hat{\beta}_0 = 39.94$ and $\hat{\beta}_1 = -0.16$ and $S^2 = 24.06$.

Model with linear quadratic terms



Parameter estimates: $\hat{\beta}_0 = 56.90$, $\hat{\beta}_1 = -0.47$ and $\hat{\beta}_2 = 0.0012$ and $S^2 = 19.13$.

There are two dual geometrical viewpoints that one may adopt:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & & x_{2q} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{(n-1)1} & x_{(n-1)2} & \dots & x_{(n-1)q} \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Row geometry: focus on the n OBSERVATIONS
- Column geometry: focus on the p EXPLANATORIES

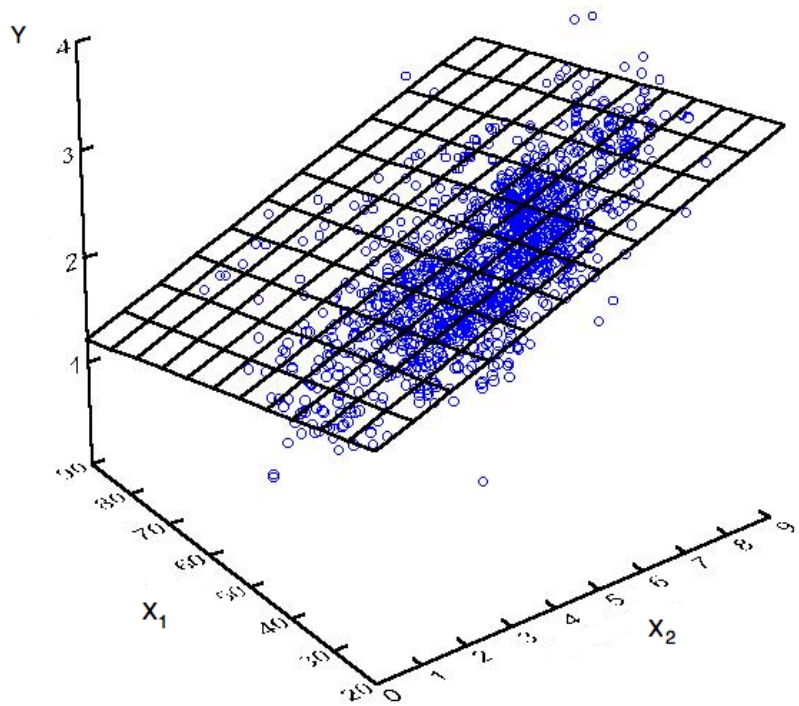
Both are useful, usually for different things:

- Row geometry useful for exploratory analysis.
- Column geometry useful for theoretical analysis.

Both geometries give useful, but different, intuitive interpretations of the least squares estimators.

Corresponds to the "scatterplot geometry" – (data space)

- n points in $\mathbb{R}^p$
- each corresponds to an observation
- least squares parameters give parametric equation for a hyperplane
- hyperplane has property that it minimizes the sum of squared vertical distances of observations from the plane itself over all possible hyperplanes



- Fitted values are vertical projections (NOT orthogonal projections!) of observations onto plane, residuals are signed vertical distances of observations from plane.

Adopt the dual perspective:

- Consider the entire vector y as a single point living in $\mathbb{R}^n$
- Then consider each variable (column) as a point also in $\mathbb{R}^n$

What is the interpretation of the p-dimensional vector $\hat{\beta}$, and the n-dimensional vectors $\hat{y}$ and e in this dual space?

Turns out there is another important plane here: the plane spanned by the variable vectors (the column vectors of X).

Recall that this is the *column space* of X, denoted by $\mathcal{M}(X)$.

$$\text{Recall:} \qquad \underbrace{\mathcal{M}(X)}_{\text{Column Space}} := \{X\gamma : \gamma \in \mathbb{R}^p\}$$

Q: What does $Y = X\beta + \varepsilon$ mean?

A: Y is [some element of $\mathcal{M}(X)$] + [Gaussian disturbance].

Any realisation y of Y will lie outside $\mathcal{M}(X)$ (almost surely). MLE estimates β by minimising

$$(y - X\beta)^\top (y - X\beta) = \|y - X\beta\|^2$$

Thus we search for a β giving the element of $\mathcal{M}(X)$ with the minimum distance from y.
Hence $\hat{y} = X\hat{\beta}$ is the projection of y onto $\mathcal{M}(X)$:

$$\hat{y} = X\hat{\beta} := \underbrace{X(X^\top X)^{-1}X^\top}_{H} y = Hy.$$

H is the *hat matrix* (puts hat on y!)

Another derivation of the MLE of β:

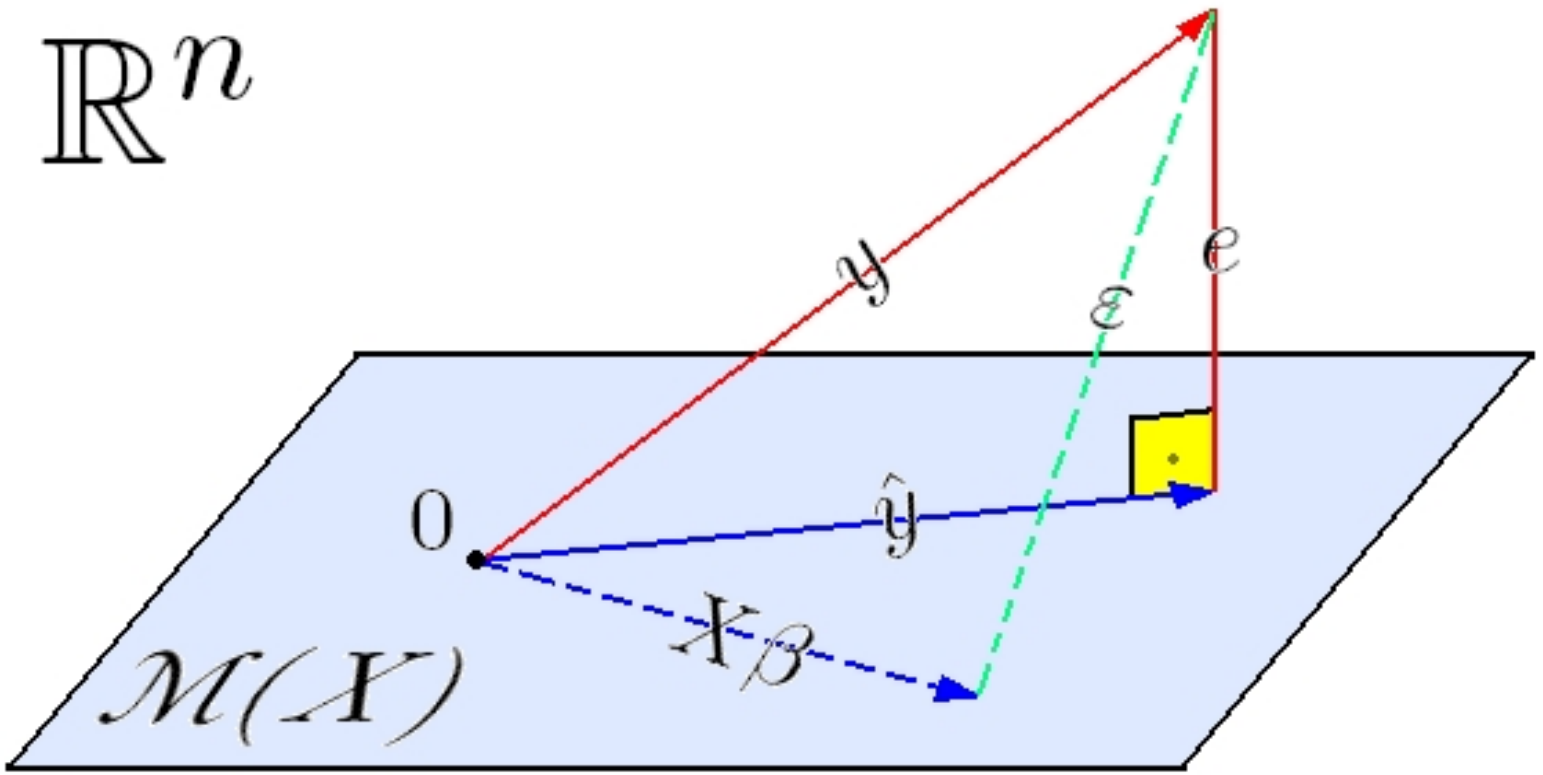
- Choose $\hat{\beta}$ to minimise $(y - X\beta)^\top (y - X\beta) = \|y - X\beta\|^2$, so

$$\hat{\beta} = \arg\min \|y - X\beta\|^2.$$

- $\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|^2 = \min_{\gamma \in \mathcal{M}(X)} \|y - \gamma\|^2$
- But the unique γ that yields $\min_{\gamma \in \mathcal{M}(X)} \|y - \gamma\|^2$ is $\gamma = Py$.
- Here P is the projection onto the column space of X, $\mathcal{M}(X)$.
- Since X is of full rank, $P = X(X^\top X)^{-1} X^\top$.
- So $\gamma = X(X^\top X)^{-1} X^\top y$
- $\hat{\beta}$ will now be the unique (since X non-singular) vector of coordinates of γ with respect to the basis of columns of X.
- So

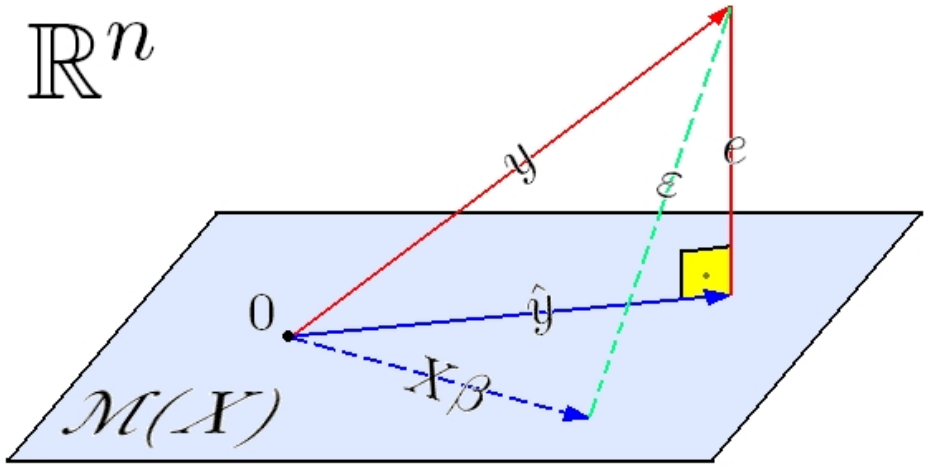
$$X\hat{\beta} = \gamma = X(X^\top X)^{-1} X^\top y,$$

  which implies that $\hat{\beta} = (X^\top X)^{-1} X^\top y$


$\mathbb{R}^n$
y
ε
e
0
$\hat{y}$
$X\beta$
$\mathcal{M}(X)$

So what is $\hat{\beta}$?

- If X columns linearly independent, they are a (non-orthogonal) basis for $\mathcal{M}$
- Hence for any $z \in \mathcal{M}(X)$, there exists a unique $\gamma \in \mathbb{R}^p$ such that $z = X\gamma$



- So γ contains coordinates of z with respect to the X-column basis
- Consequently, $\hat{\beta}$ contains coordinates of $\hat{y}$ with respect to the X-column basis
- But $\hat{y} = Hy = X\underbrace{(X^\top X)^{-1}X^\top y}_{u} = Xu$, so u is the unique vector that gives coordinates of y with respect to the X-column basis
- Hence we must have $\hat{\beta} = u = (X^\top X)^{-1}X^\top y$

Facts:

1. $e = (I - H)y = (I - H)\varepsilon$.
2. $\hat{y}$ and e are orthogonal, i.e. $\hat{y}^\top e = 0$
3. Pythagoras: $y^\top y = \hat{y}^\top \hat{y} + e^\top e = y^\top H y + \varepsilon^\top (I - H)\varepsilon$

Derivation:

1. $e = y - X\hat{\beta} = y - Hy = (I - H)y = (I - H)(X\beta + \varepsilon) = (I - H)X\beta + (I - H)\varepsilon = (I - H)\varepsilon$
2. $e = y - \hat{y} = (I - H)y \implies \hat{y}^\top e = y^\top H^\top (I - H)y = 0$
3. $y^\top y = (Hy + (I - H)y)^\top (Hy + (I - H)y) = \hat{y}^\top \hat{y} + e^\top e + \underbrace{2yH(I - H)y}_{=0}.$

Assume slightly different model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \frac{\varepsilon_i}{\sqrt{w_i}}, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma^2), \quad w_i > 0$$

$$\Updownarrow$$

$$Y_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}, \frac{\sigma^2}{w_i}\right).$$

With the w_j *known* weights (example: each Y_j is an average of w_j measurements).

Arises often in practice (e.g., in sample surveys), but also arises in theory.

Transformation:

$$y' = W^{1/2} y, \quad X' = W^{1/2} X$$

with

$$W_{n \times n} = \text{diag}(w_1, \ldots, w_n)$$

Leads to usual scenario. In this notation we obtain:

$$\begin{aligned} \hat{\beta} &= [(X')^\top X']^{-1} (X')^\top y' \\ &= (X^\top W X)^{-1} X^\top W y \end{aligned}$$

Similarly:

$$S^2 = \frac{1}{n-p} y^\top \left[W - WX(X^\top WX)^{-1} X^\top W \right] y$$

Distribution Theory of Least Squares

Gaussian Linear Model:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

We have derived the estimators:

- $\hat{\beta} = (X^\top X)^{-1} X^\top y$
- $\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\top (y - X\hat{\beta}) = \frac{1}{n}\|\hat{y} - y\|^2$
- $S^2 = \frac{1}{n-p}\|\hat{y} - y\|^2$

We need to study the distribution of these estimators for the purpose of:

- Understanding their precision
- Building confidence intervals
- Testing hypotheses
- Comparing them to other candidate estimators
- ...

Theorem

Let $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$ *with* $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$ *and assume that* X *has full rank* $p < n$*. Then,*

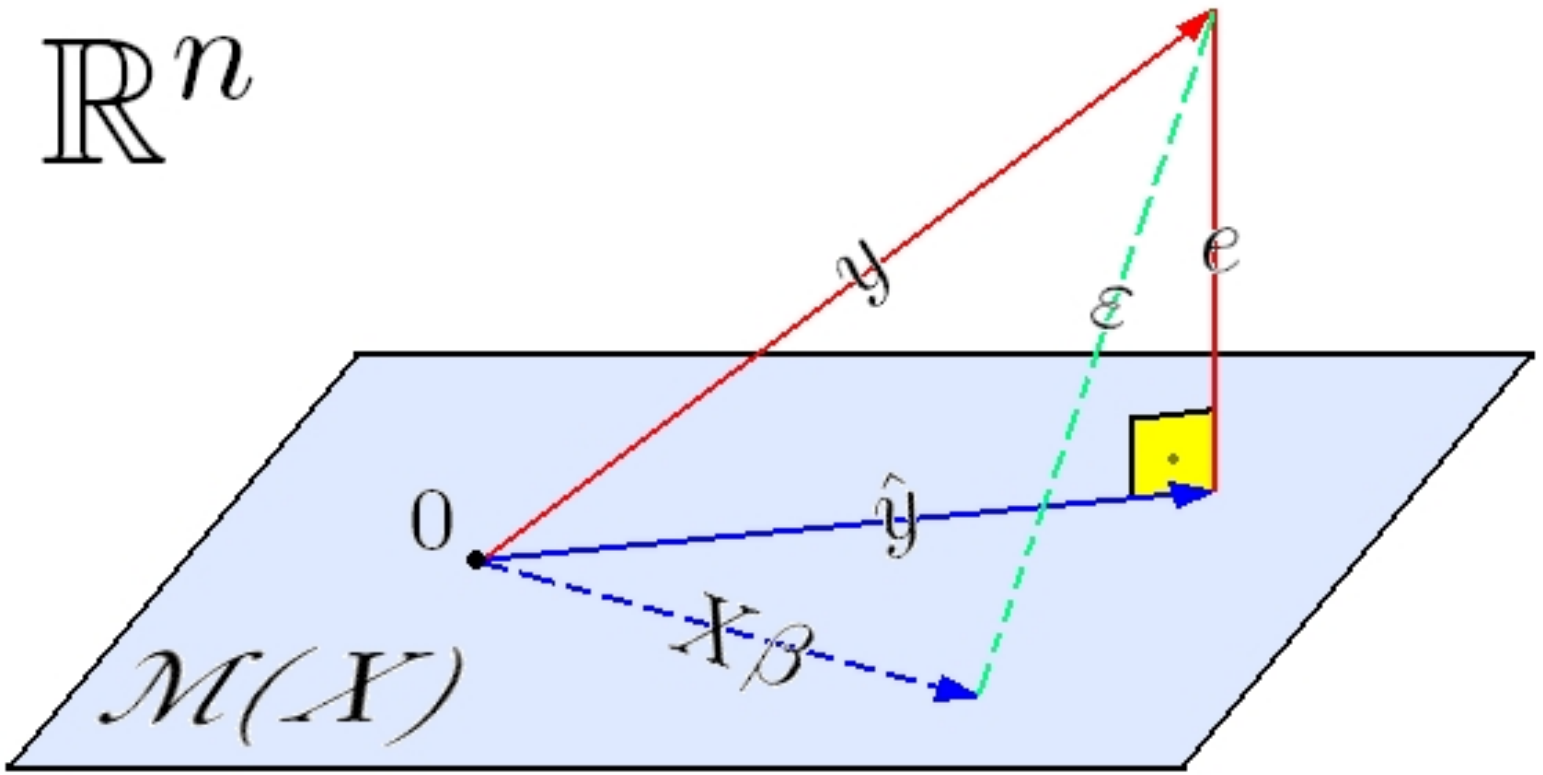
1. $\hat{\beta} \sim \mathcal{N}_p\{\beta, \sigma^2 (X^\top X)^{-1}\}$;
2. *the random variables* $\hat{\beta}$ *and* S^2 *are independent; and*
3. $\frac{n-p}{\sigma^2} S^2 \sim \chi^2_{n-p}$*, where* χ^2_ν *denotes the chi-square distribution with* ν *degrees of freedom.*

Corollary

Let $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$ *with* $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$*. The statistic* Hy *is sufficient for the parameter* β*. If* X *has full rank* $p < n$*, then* $\hat{\beta}$ *is also sufficient for* β*.*

Corollary

S^2 *is unbiased whereas* $\hat{\sigma}^2$ *is biased (so we prefer* S^2*).*


$\mathbb{R}^n$
y
ε
e
0
$\hat{y}$
$X\beta$
$\mathcal{M}(X)$

Proof of the Theorem.

1. Recall our results for linear transformations of Gaussian variables:

$$\left.\begin{array}{l} \hat{\beta} = (X^\top X)^{-1} X^\top Y \\ Y \sim \mathcal{N}_n(X\beta, \sigma^2 I) \end{array}\right\} \Longrightarrow \hat{\beta} \sim \mathcal{N}_p\{\beta, \sigma^2 (X^\top X)^{-1}\}$$

2. If e is independent of $\hat{y} = X\hat{\beta}$, then $S^2 = e^\top e/(n-p)$ will be independent of $\hat{\beta}$ (why?). Now notice that:

- $e = (I - H)y$
- $\hat{y} = Hy$
- $y \sim \mathcal{N}(X\beta, \sigma^2 I)$

Therefore, from the properties of the Gaussian distribution e is independent of $\hat{y}$ since $(I - H)(\sigma^2 I)H = \sigma^2(I - H)H = 0$, by idempotency of H.

proof cont'd.

3. For the last part recall that

$$e = (I - H)\varepsilon \implies (n-p)S^2 = (n-p)\frac{e^\top e}{n-p} = \varepsilon^\top (I-H)\varepsilon$$

by idempotency of H. But recall that $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ so $\sigma^{-1}\varepsilon \sim \mathcal{N}_n(0, I_n)$. Therefore, by the properties of normal quadratic forms (slide 40),

$$\frac{(n-p)}{\sigma^2}S^2 = (\sigma^{-1}\varepsilon)^\top (I-H)(\sigma^{-1}\varepsilon) \sim \chi^2_{n-p}.$$

□

Proof of the first Corollary.

Write $y = Hy + (I - H)y = \hat{y} + e$.

If we can show that the conditional distribution of the $2n$-dimensional vector $W = (\hat{y}, e)^\top$ given $\hat{y}$ does not depend on β, then we will also know that the conditional distribution of $y = \hat{y} + e$ given $\hat{y}$ does not depend on β either, proving the proposition.

But we have proven that $\hat{y}$ is independent of e. Therefore, conditional on $\hat{y}$, e always has the same distribution $\mathcal{N}(0, (I - H)\sigma^2)$. It follows that, conditional on $\hat{y}$, the vector W has a distribution whose first n coordinates equal $\hat{y}$ almost surely, and whose last n coordinates are $\mathcal{N}(0, (I - H)\sigma^2)$. Neither of those two depend on β, and the proof is complete.

When X has full rank, $\hat{\beta}$ is a 1-1 function of Hy, and is also sufficient for β.

□

Proof of the second Corollary.

Recall that if $Q \sim \chi^2_d$, then $\mathbb{E}[Q] = d$. □

How to construct $1-\alpha$ CI for a linear combination of the parameters, $c^\top \beta$?

- Have $c^\top \hat{\beta} \sim \mathcal{N}_1(c^\top \beta, \sigma^2 c^\top (X^\top X)^{-1} c) = \mathcal{N}_1(c^\top \beta, \sigma^2 \delta)$
- Therefore $Q = (c^\top \hat{\beta} - c^\top \beta)/(\sigma\sqrt{\delta}) \sim \mathcal{N}_1(0,1)$
- Hence $Q^2 \sim \chi^2_1$
- and Q^2 is independent of S^2 (since $\hat{\beta}$ is independent of S^2)
- while $\frac{n-p}{\sigma^2} S^2 \sim \chi^2_{n-p}$.

In conclusion:

$$\frac{\frac{Q^2}{1}}{\frac{\frac{(n-p)}{\sigma^2} S^2}{n-p}} \sim F_{1,n-p} \Rightarrow \frac{\frac{(c^\top \hat{\beta} - c^\top \beta)^2}{\sigma^2 \delta}}{\frac{S^2}{\sigma^2}} = \left(\frac{c^\top \hat{\beta} - c^\top \beta}{\sqrt{S^2 c^\top (X^\top X)^{-1} c}} \right)^2 \sim F_{1,n-p}$$

- But for real W, $W^2 \sim F_{1,n-p} \iff W \sim t_{n-p}$, so base CI on:

$$\frac{c^\top \hat{\beta} - c^\top \beta}{\sqrt{S^2 c^\top (X^\top X)^{-1} c}} \sim t_{n-p}$$

- We obtain $(1-\alpha)\times 100\%$ CI:

$$c^\top \hat{\beta} \pm t_{n-p}(1-\alpha/2)\sqrt{S^2 c^\top (X^\top X)^{-1} c}.$$

- What about a $(1-\alpha)$ CI for β_r? (rth coordinate)
- Let $c_r = (0, 0, \ldots, 0, \underbrace{1}_{r^{th} \text{ position}}, 0, \ldots, 0)$
- Then $\beta_r = c^\top \beta$
- Therefore, base CI on

$$\frac{c_r^\top \hat{\beta} - c_r^\top \beta}{\sqrt{S^2 c_r^\top (X^\top X)^{-1} c_r}} = \frac{\hat{\beta}_r - \beta_r}{\sqrt{S^2 v_{r,r}}} \sim t_{n-p},$$

  where $v_{r,s}$ is the r,s element of $(X^\top X)^{-1}$.
- Obtain $(1-\alpha)\times 100\%$ CI:

$$\hat{\beta} \pm t_{n-p}(1-\alpha/2)\sqrt{S^2 v_{rr}}.$$

- Suppose we want to predict the value of y_+ for an $x_+ \in \mathbb{R}^p$
- Our model predicts y_+ by $x_+^\top \hat{\beta}$.
- But $y_+ = x_+^\top \beta + \varepsilon_+$ so a prediction interval is DIFFERENT from an interval for a linear combination $c^\top \beta$ (extra uncertainty due to ε_+):
  - $\mathbb{E}[x_+^\top \hat{\beta} + \varepsilon_+] = x_+^\top \beta$
  - $\text{var}[x_+^\top \hat{\beta} + \varepsilon_+] = \text{var}[x_+^\top \hat{\beta}] + \text{var}[\varepsilon_+] = \sigma^2[x_+^\top (X^\top X)^{-1} x_+ + 1]$
- Base prediction interval on:

$$\frac{x_+^\top \hat{\beta} - y_+}{\sqrt{S^2\{1 + x_+^\top (X^\top X)^{-1} x_+\}}} \sim t_{n-p}.$$

- Obtain $(1-\alpha)$ prediction interval:

$$x_+^\top \hat{\beta} \pm t_{n-p}(1-\alpha/2)\sqrt{S^2\{1 + x_+^\top (X^\top X)^{-1} x_+\}}.$$

R^2 is a *measure of fit* of the model to the data.

- We are trying to best approximate y through an element of the column-space of X.
- How successful are we? Squared error is $e^\top e$.
- How large is this, relative to data variation? Look at

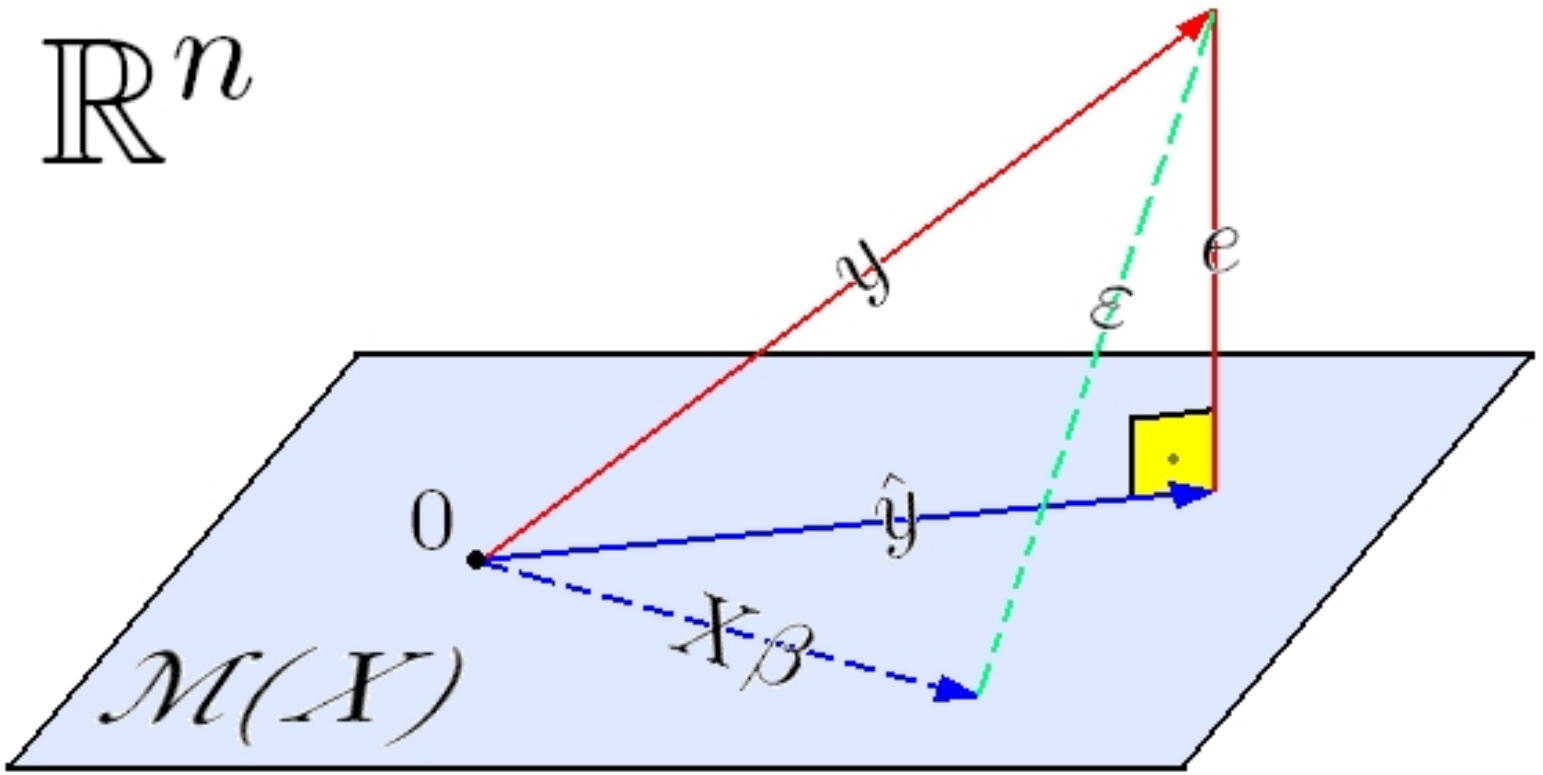
$$\frac{\|e\|^2}{\|y\|^2} = \frac{e^\top e}{y^\top y} = \frac{y^\top (I - H) y}{y^\top y} = 1 - \frac{\hat{y}^\top \hat{y}}{y^\top y}$$

- Define

$$R_0^2 = \frac{\hat{y}^\top \hat{y}}{y^\top y} = \frac{\|\hat{y}\|^2}{\|y\|^2}$$

- Note that $0 \leq R_0^2 \leq 1$

Interpretation: what proportion of the squared norm of y does our fitted value $\hat{y}$ explain?


$\mathbb{R}^n$
y
ε
e
0
$\hat{y}$
$X\beta$
$\mathcal{M}(X)$

"**Centred (in fact, usual)** R^2". Compares empirical variance of $\hat{y}$ to empirical variance of y, instead of the empirical norms. In other words:

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^n y_i^2 - n\bar{y}^2}.$$

(note that $\frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n (y_i - e_i) = \bar{y}$ because $e \perp \mathbf{1}$ (recall that $\mathbf{1}$ is the vector of 1's = first column of design matrix X) so $\sum_i e_i = 0$.

Note that

$$R^2 = \frac{\|\hat{y}\|^2 - \|\bar{y}\mathbf{1}\|^2}{\|y\|^2 - \|\bar{y}\mathbf{1}\|^2}.$$

- R_0^2 mathematically more natural (does not treat first column of X as special).
- R^2 statistically more relevant (expresses variance—the first column of X usually *is* special, in statistical terms!).
- R_0^2 and R^2 may differ a lot when $\bar{y}$ large.

Geometrical interpretation of R^2: project y and $\hat{y}$ on orthogonal complement of $\mathbf{1}$, then compare the norms (of the projections):

- $\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top y = \mathbf{1}n^{-1}\sum_{i=1}^n y_i = \mathbf{1}\bar{y}$.
- $\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top \hat{y} = \mathbf{1}n^{-1}\sum_{i=1}^n \hat{y}_i = \mathbf{1}\bar{y}$.

So

$$R^2 = \frac{\|\hat{y}\|^2 - \|\bar{y}\mathbf{1}\|^2}{\|y\|^2 - \|\bar{y}\mathbf{1}\|^2} = \frac{\|(I - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1})\hat{y}\|^2}{\|(I - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1})y\|^2}$$

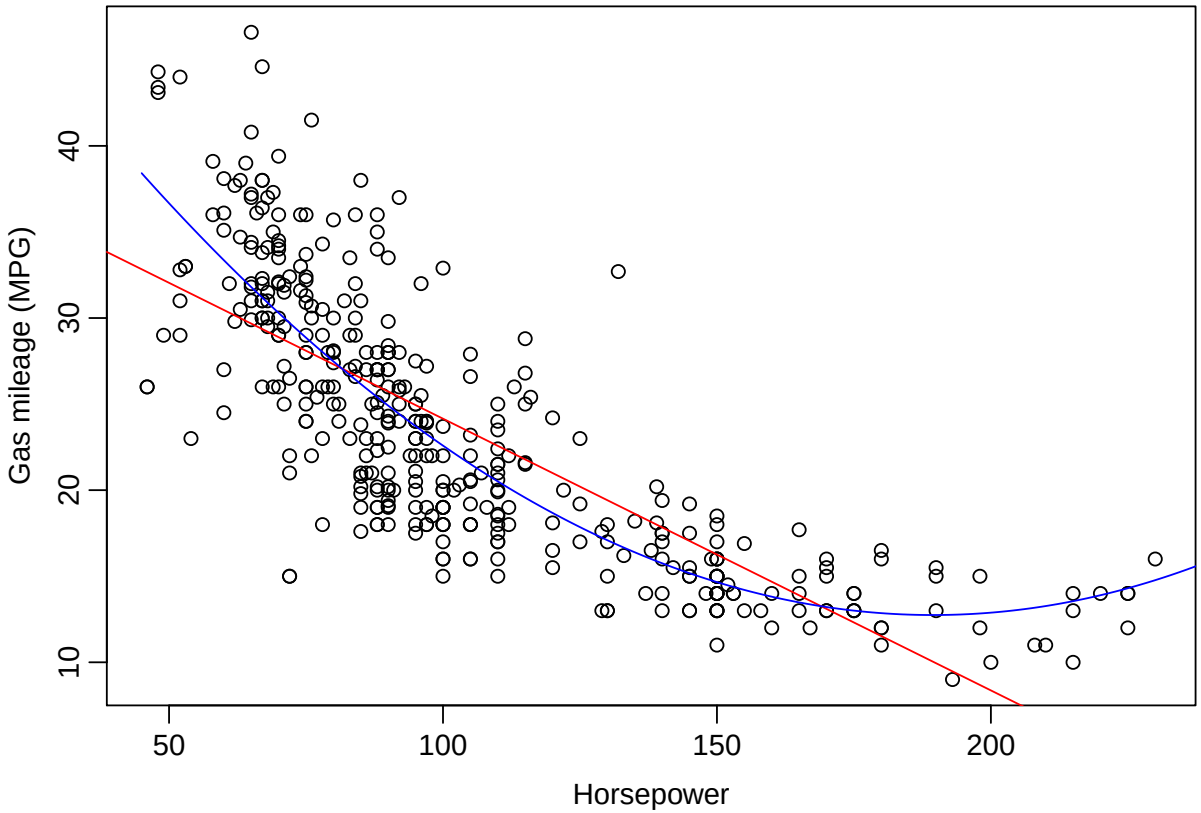
Intuition: Should not take into account the part of $\|y\|$ that is explained by a constant, we only want to see the effect of the explanatory variables.

NOTE: Statistical packages (e.g., `R`) provide R^2 (and/or R^2_a, see below), not R^2_0.

Exercise: Show that $R^2 = [\text{corr}(\{\hat{y}_i\}_{i=1}^n, \{y_i\}_{i=1}^n)]^2$.
Exercise: Show that $R^2 \leq R^2_0$.

R^2 coefficients for the linear and quadratic models:

	R_0^2	R^2
linear	0.96	0.61
quadratic	0.97	0.69



The adjusted R^2 takes into account the number of variables employed. It is defined as:

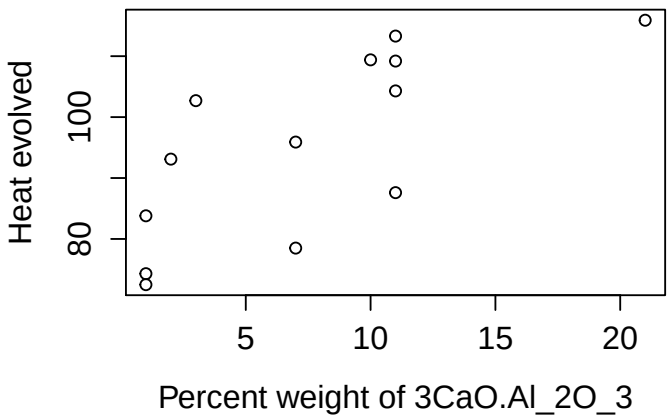
$$R_a^2 = R^2 - (1 - R^2)\frac{n-1}{n-p}.$$

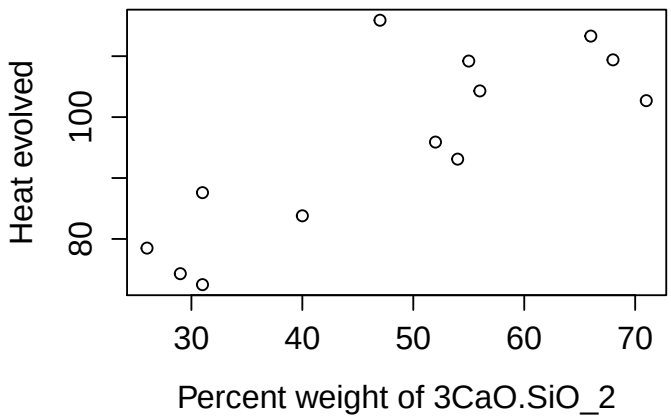
Corrects for the fact that we can always increase R^2 by adding variables. One can also correct the un-centred R_0^2 by evaluating

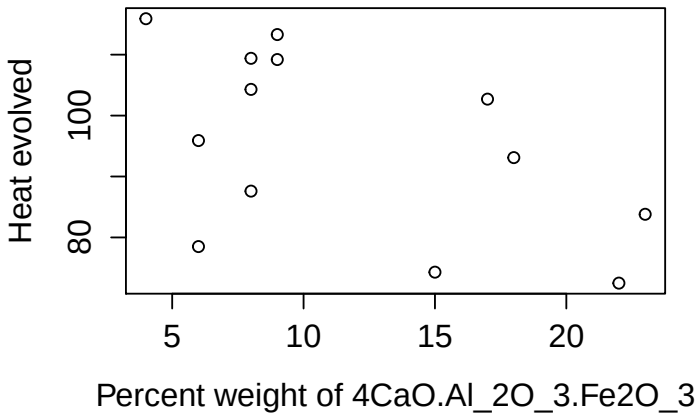
$$R_0^2 - (1 - R_0^2)\frac{n}{n-p}.$$

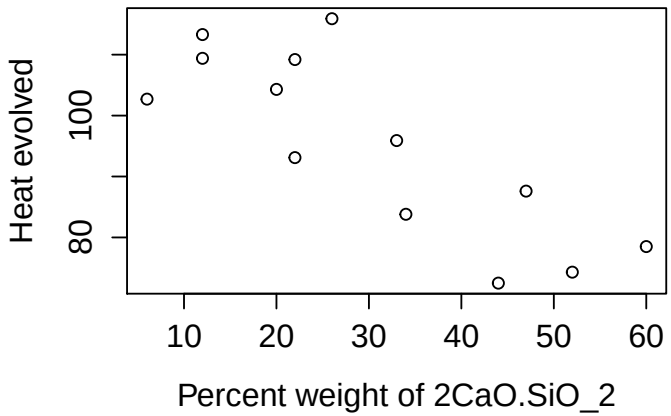
Case	$3\,CaO.Al_2O_3$	$3\,CaO.SiO_2$	$4\,Cao.Al_2O_3.Fe_2O_3$	$2\,CaO.SiO_2$	Heat
1	7.00	26.00	6.00	60.00	78.50
2	1.00	29.00	15.00	52.00	74.30
3	11.00	56.00	8.00	20.00	104.30
4	11.00	31.00	8.00	47.00	87.60
5	7.00	52.00	6.00	33.00	95.90
6	11.00	55.00	9.00	22.00	109.20
7	3.00	71.00	17.00	6.00	102.70
8	1.00	31.00	22.00	44.00	72.50
9	2.00	54.00	18.00	22.00	93.10
10	21.00	47.00	4.00	26.00	115.90
11	1.00	40.00	23.00	34.00	83.80
12	11.00	66.00	9.00	12.00	113.30
13	10.00	68.00	8.00	12.00	109.40

Example: Cement Heat Evolution









```
> cement.lm<-lm(y~1+x1+x2+x3+x4,data=cement)
> summary(cement.lm)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.4054	70.0710	0.89	0.3991
x1	1.5511	0.7448	2.08	0.0708
x2	0.5102	0.7238	0.70	0.5009
x3	0.1019	0.7547	0.14	0.8959
x4	−0.1441	0.7091	−0.20	0.8441

```
Residual standard error:  2.446 on 8 degrees of freedom
R-Squared:  0.9824
```

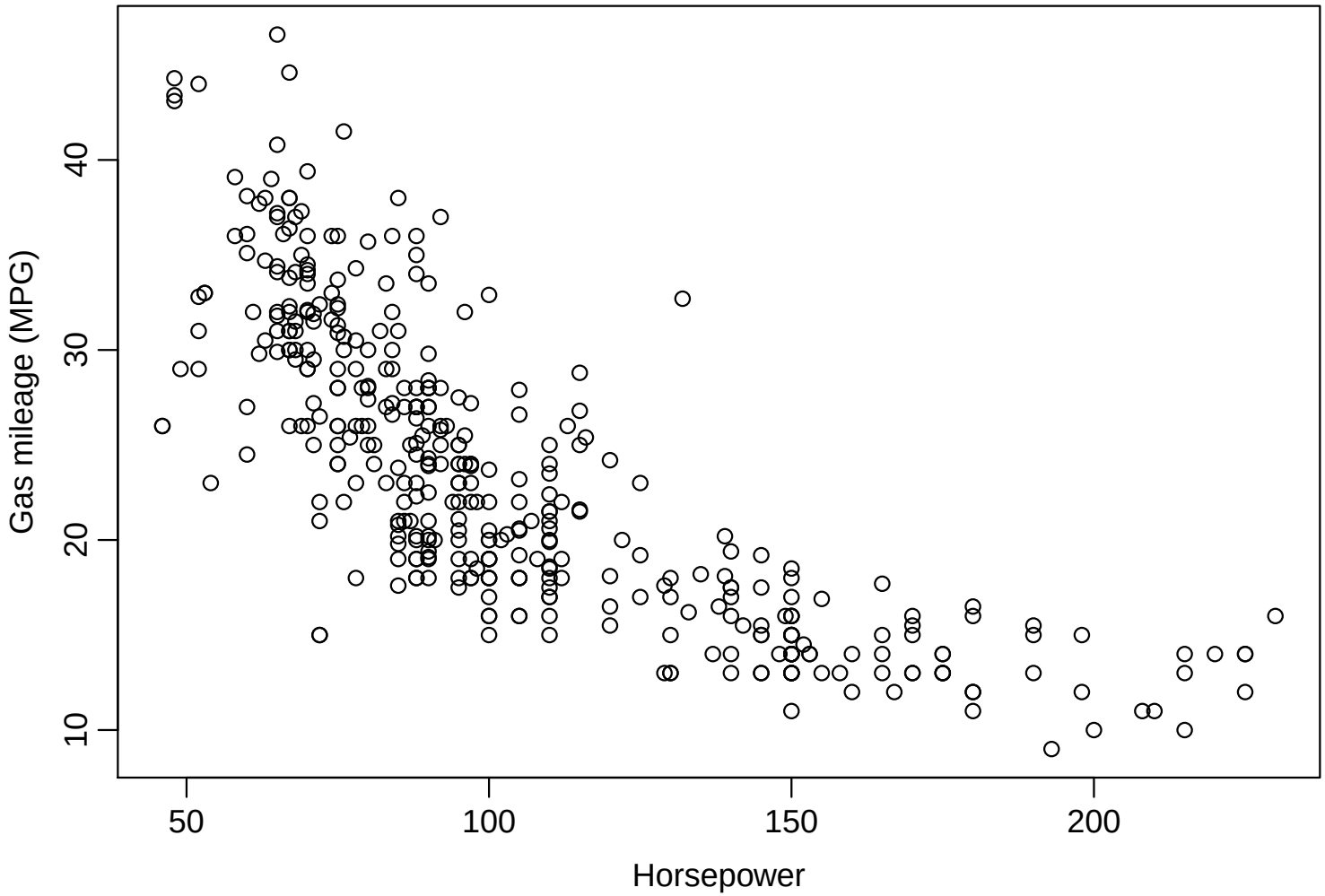
```
> x.plus
[1] 25 25 25 25
predict(cement.lm,x.plus,interval="confidence",
se.fit=T,level=0.95)
```

Fit	Lower	Upper
112.8	97.5	128.2

```
predict(cement.lm,x.plus,interval="prediction",
se.fit=T,level=0.95)
```

Fit	Lower	Upper
112.8	96.5	129.2

Horsepower and MPG of cars



```
> auto.lm <- lm(mpg~1+horsepower+I(horsepower^2),data=Auto)
> summary(auto.lm)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	56.9000	1.8004	31.60	$< 2 \times 10^{-16}$
horsepower	−0.4662	0.0311	−14.98	$< 2 \times 10^{-16}$
I(horsepower^2)	0.0012	0.0001	10.08	$< 2 \times 10^{-16}$

```
Residual standard error:  4.374 on 389 degrees of freedom
R-Squared:  0.6876
```

```
> x.plus
horsepower
       120
> predict(auto.lm, x.plus, interval="confidence",
se.fit=T, level=0.95)
```

Fit	Lower	Upper
18.68	18.03	19.33

```
> predict(auto.lm, x.plus, interval="prediction",
se.fit=T, level=0.95)
```

Fit	Lower	Upper
18.68	10.05	27.30

Gauss-Markov & Optimal Estimation

Q: Geometry suggests that the LSE $\hat{\beta}$ is a sensible estimator. But is it the best we can come up with?

A: Yes, $\hat{\beta}$ is the *unique minimum variance unbiased estimator* of β.

(To be seen in Statistical Theory course, since $\hat{\beta}$ is sufficient *and* complete)

Thus, in the Gaussian Linear model, the LSE are the best we can do as far as unbiased estimators go.

(actually can show S^2 is optimal unbiased estimator of σ^2, by similar arguments)

The crucial assumption so far was:

- **Normality**: $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$

What if we drop this strong assumption and assume something weaker?

- **Uncorrelatedness**: $\mathbb{E}[\varepsilon] = 0$ & $\text{var}[\varepsilon] = \sigma^2 I$

(notice we do not assume any particular distribution.)

How well do our LSE estimators perform in this case?

(note that in this setup the observations may not be independent — uncorrelatedness implies independence only in the Gaussian case.)

For a start, we retain unbiasedness:

Lemma

If we only assume both

$$\mathbb{E}[\epsilon] = 0 \quad var[\epsilon] = \sigma^2 I$$

instead of

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

then the following remain true:

1. $\mathbb{E}[\hat{\beta}] = \beta$;
2. $Var[\hat{\beta}] = \sigma^2 (X^\top X)^{-1}$;
3. $\mathbb{E}[S^2] = \sigma^2$.

But what about optimality properties?

Theorem

Let $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$, *with* $p < n$, X *having rank* p, *and*

- $\mathbb{E}[\varepsilon] = 0$,
- $var[\varepsilon] = \sigma^2 I$.

Then, $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ *is the best linear unbiased estimator of* β, *that is, for any linear unbiased estimator* $\tilde{\beta}$ *of* β, *it holds that*

$$var(\tilde{\beta}) - var(\hat{\beta}) \succeq 0.$$

Proof.

Let $\tilde{\beta}$ be linear and unbiased, in other words: $\begin{cases} \tilde{\beta} = AY, & \text{for some } A_{p\times n}, \\ \mathbb{E}[\tilde{\beta}] = \beta, & \text{for all } \beta \in \mathbb{R}^p. \end{cases}$

These two properties combine to yield,

$$\beta = \mathbb{E}[\tilde{\beta}] = \mathbb{E}[AY] = \mathbb{E}[AX\beta + A\varepsilon] = AX\beta, \quad \beta \in \mathbb{R}^p$$

$$\implies (AX - I)\beta = 0, \forall \beta \in \mathbb{R}^p.$$

We conclude that the null space of $(AX - I)$ is the entire $\mathbb{R}^p$, and so $AX = I$.

$$\begin{aligned} \text{var}[\tilde{\beta}] - \text{var}[\hat{\beta}] &= A\sigma^2 I A^\top - \sigma^2 (X^\top X)^{-1} \\ &= \sigma^2 \{AA^\top - AX(X^\top X)^{-1} X^\top A^\top\} \\ &= \sigma^2 A(I - H)A^\top \\ &= \sigma^2 A(I - H)(I - H)^\top A^\top \\ &\succeq 0. \end{aligned}$$

□

If $\mathbb{E}[\varepsilon] = 0$ and $\text{cov}[\varepsilon] = \sigma^2 I$

$\hookrightarrow$Gauss-Markov says $\hat{\beta}$ optimal linear unbiased estimator, regardless of whether or not ε is Gaussian.

Question: *What can we say about the distribution of $\hat{\beta}$ when $\text{cov}(\varepsilon) = \sigma^2 I$, but ε is not necessarily Gaussian?*

Note that we can always write

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

- Since there is a huge variety of candidate distributions for ε that would be compatible with the property $\text{cov}(\varepsilon) = \sigma^2 I$, we cannot say very much about the exact distribution of $\hat{\beta} - \beta = (X^\top X)^{-1} X^\top \varepsilon$.
- Can we at least hope to say something about this distribution asymptotically, as the sample becomes large?
- When $Y_i = \mu + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, ..., n$, we have

$$\hat{\beta} - \beta = (\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1}^\top \varepsilon = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i$$

and the usual CLT applies as $n \to \infty$. Can we generalise this?

Large sample $\iff$ increasing number of observations.

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

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

For $p \geq 1$ fixed, let $\{X_n\}_{n \geq p}$ be a sequence of $n \times p$ design matrices and $Y_n = X_n \beta + \varepsilon_n$. If

1. *X_n is of full rank p for all $n \geq p$*
2. $\max_{1 \leq i \leq n} [x_i^\top (X_n^\top X_n)^{-1} x_i] \overset{n \to \infty}{\longrightarrow} 0$, *[known as Noether's condition[a]]*
   (where $x_i^\top$ is the ith row of X_n)
3. *ε_n is a zero mean n-vector with i.i.d. coordinates of variance σ^2,*

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

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

[a] Gottfried Noether (not Emmy Noether), Ann. Math. Stat., 1949.

Theorem's conclusion can be interpreted as:

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

- i.e. distribution of $\hat{\beta}$ gradually becomes what it would be if ε were Gaussian
- . . . provided design matrix X satisfies Noether's condition (2).
- This 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* $n \to \infty$

  Because $x_i^\top (X^\top X)^{-1} x_i = (e_i^\top X)(X^\top X)^{-1}(e_i^\top X)^\top = e_i^\top H e_i = h_{ii}$ where e_i is the ith canonical basis vector for $\mathbb{R}^n$.
- Note that 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

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

Here, $p = 2$. Can show that

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

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

$$h_{jj}(n) = \frac{1}{n} + \frac{\{j - (n+1)/2\}^2}{(n^3 - n)/12}$$

so $$\max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) = \frac{1}{n} + \frac{3(n-1)}{n^2(n-1)} \overset{n \to \infty}{\longrightarrow} 0.$$

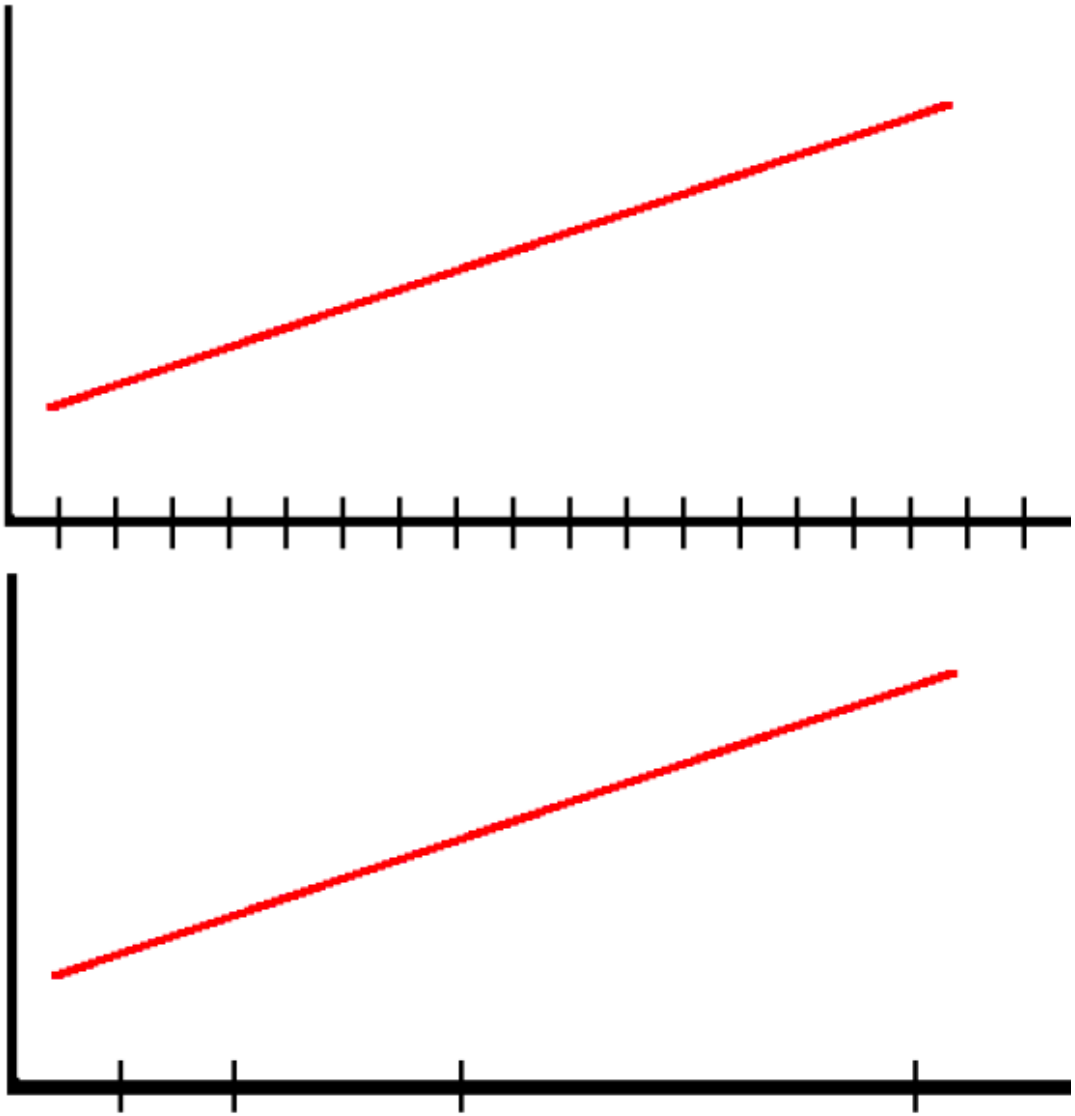
- Now consider $t_i = 2^i$ (grid points spread apart as n grows).
  The centre of mass and sum of squares of the grid points is now

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

  and so

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

Large Sample Distribution of $\hat{\beta}$



Proof.

Recall that $\hat{\beta}_n - \beta = (X_n^\top X_n)^{-1} X_n^\top \varepsilon_n$. We will show that for any unit vector u,

$$u^\top (X_n^\top X_n)^{-1/2} X_n^\top \varepsilon_n \xrightarrow{d} N(0, \sigma^2),$$

and then the theorem will be proven by the Cramér-Wold device[a]. Now notice that

$$u^\top (X_n^\top X_n)^{-1/2} X_n^\top \varepsilon_n = \gamma_n^\top \varepsilon_n$$

where:

1. $\gamma_n = (\gamma_{n,1}, \ldots, \gamma_{n,n})^\top = \left(u^\top (X_n^\top X_n)^{-1/2} x_1, \ldots, u^\top (X_n^\top X_n)^{-1/2} x_n\right)^\top$
2. $\gamma_{n,i}^2 \leq \|u\|^2 \left\|(X_n^\top X_n)^{-1/2} x_i\right\|^2 = x_i^\top (X_n^\top X_n)^{-1} x_i$ (Cauchy-Schwarz)
3. $\gamma_n^\top \gamma_n = u^\top (X_n^\top X_n)^{-1/2} (X_n^\top X_n) (X_n^\top X_n)^{-1/2} u = 1.$

Consequently, the result follows from the weighted sum CLT upon noticing:

$$\max_{1 \leq i \leq n} \gamma_{n,i}^2 \Big/ \sum_{k=1}^n \gamma_{n,k}^2 \leq \max_{1 \leq i \leq n} x_i^\top (X_n^\top X_n)^{-1} x_i \to 0$$

---

[a] Cramér-Wold: $\xi_n \xrightarrow{d} \xi$ in $\mathbb{R}^d$ if and only if $u^\top \xi \xrightarrow{d} u^\top \xi$ in $\mathbb{R}$ for all unit vectors u.

Diagnostics

Four basic assumptions inherent in the Gaussian linear regression model:

- **Linearity**: $\mathbb{E}[Y]$ is linear in X.
- **Homoskedasticity**: $\text{var}[\varepsilon_j] = \sigma^2$ for all $j = 1, \ldots, n$.
- **Gaussian Distribution**: errors are Normally distributed.
- **Independent Errors**: ε_i independent of ε_j for $i \neq j$.

When one of these assumptions fails clearly, then Gaussian linear regression is inappropriate as a model for the data.

Isolated problems, such as outliers and influential observations also deserve investigation. They *may or may not* decisively affect model validity.

Scientific reasoning: impossible to *validate* model assumptions.

Cannot *prove* that the assumptions hold. Can only provide evidence in favour (or against!) them.

Strategy:

- Find implications of each assumption that we can check graphically (mostly concerning residuals).
- Construct appropriate plots and assess them (requires experience).

"Magical Thinking": Beware of overinterpreting plots!

Residuals e: Basic tool for checking assumptions.

$$\underline{\text{Recall}}: \; e = y - \hat{y} = y - X\hat{\beta} = (I - H)y = (I - H)\varepsilon$$

Intuition: the residuals represent the aspects of y that cannot be explained by the columns of X.

Since $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$, if the model is correct we should have $e \sim \mathcal{N}_n\{0, \sigma^2(I - H)\}$.

$$\text{So if assumptions hold} \rightarrow \begin{cases} e_i \sim \mathcal{N}\{0, \sigma^2(1 - h_{ii})\} \\ \text{cov}(e_i, e_j) = -\sigma^2 h_{ij} \end{cases}$$

Note the residuals are correlated, and that they have unequal variances. Define the *standardised residuals*:

$$r_i := \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n.$$

These are still correlated but have variance ≈ 1.
(can decorrelate by $U^\top e$, where $H = U\Lambda U^\top$) – why?

Is $\mathbb{E}[Y]$ entirely specified as linear functional of X? Did we leave variables out? Is it also a linear functional of non-linear transformations of X-columns?
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, by construction of $e = (I - H)y$ we have

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

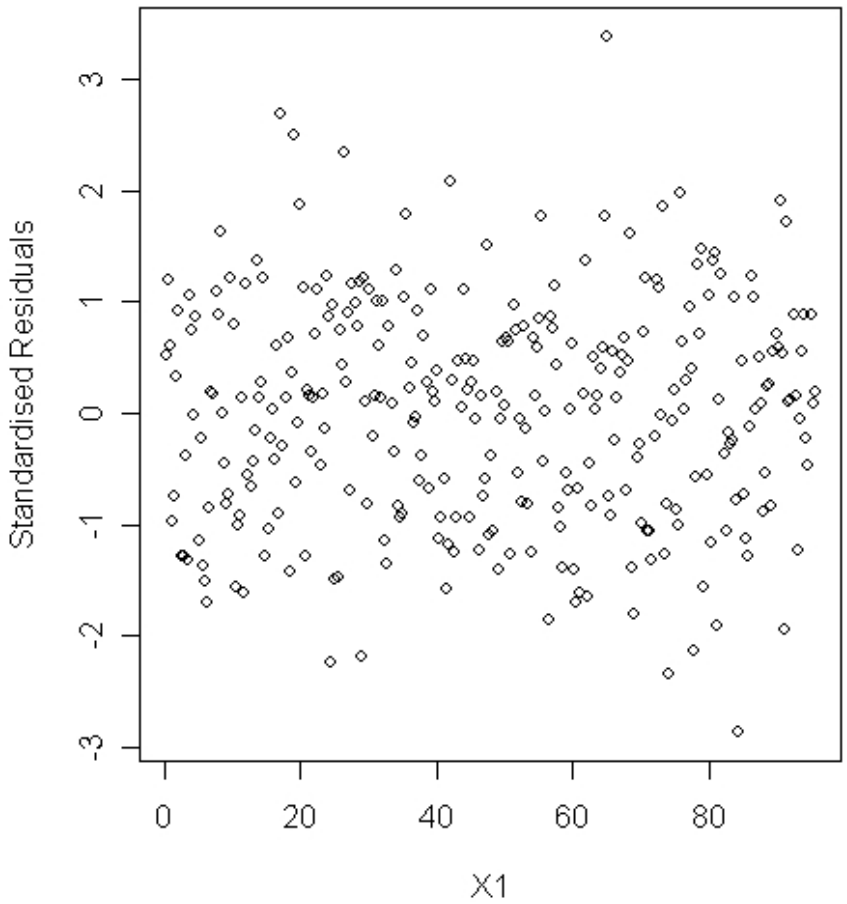
Hence, no correlation **will** appear between explanatory variables and residuals.

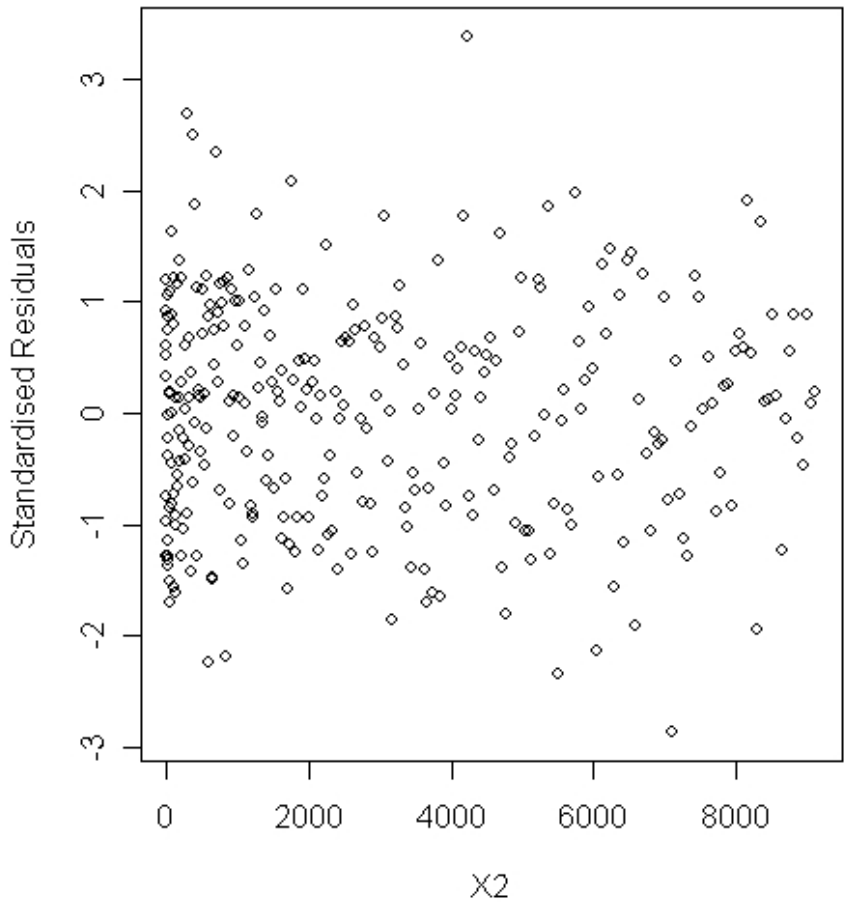
- Plot stand. residuals r against each explanatory variable (columns of X).
  - ↪ No systematic (non-linear) patterns should appear in these plots. A systematic pattern would suggest incorrect dependence of the response on the particular explanatory (e.g. need to add a transformation of that explanatory as an additional variable).

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

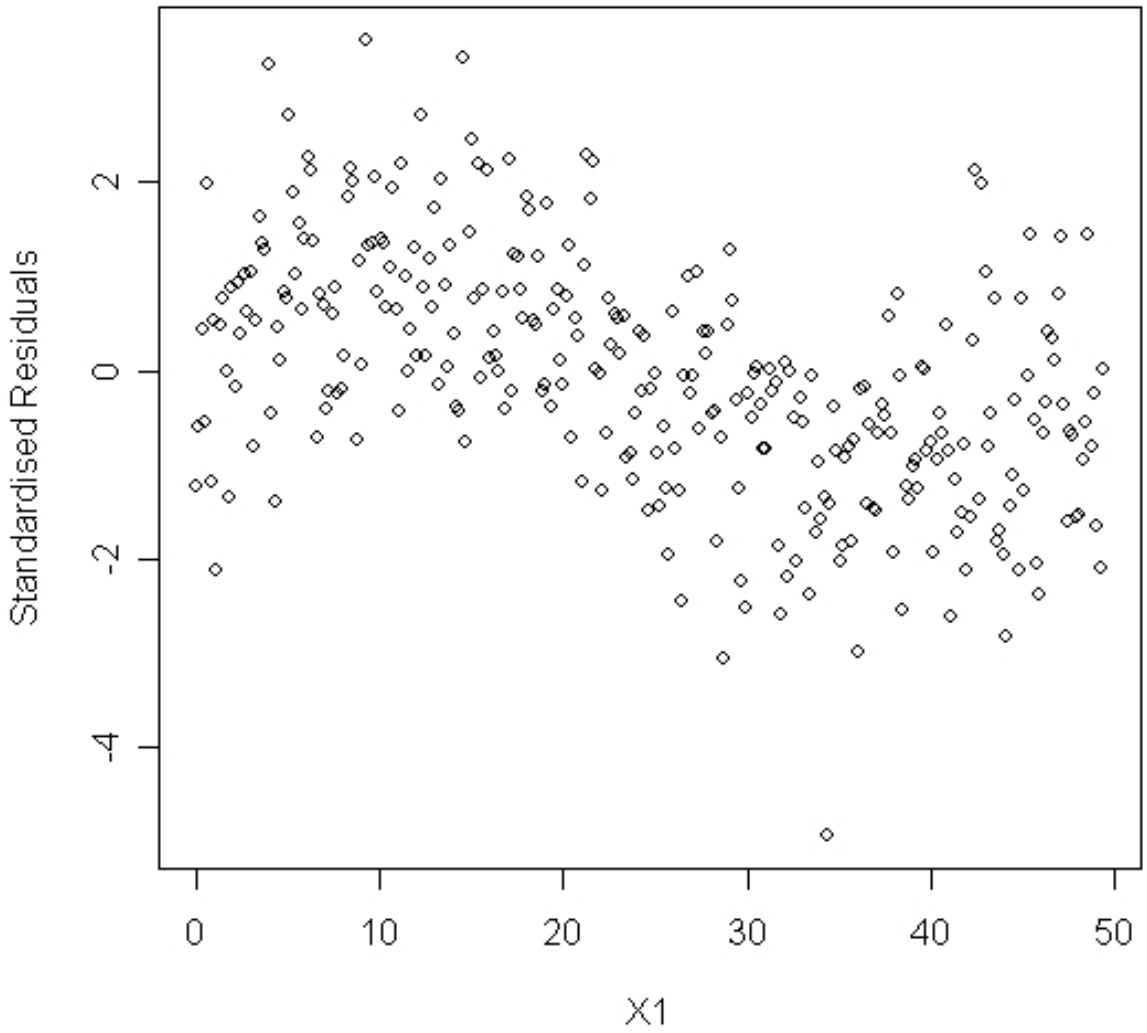
- Plot standardised residuals r against explanatories left out of the model.
  - ↪ No systematic patterns should appear in these plots. A systematic pattern suggests that we have left out an explanatory variable (or transformation thereof) that should have been included.

Linearity OK

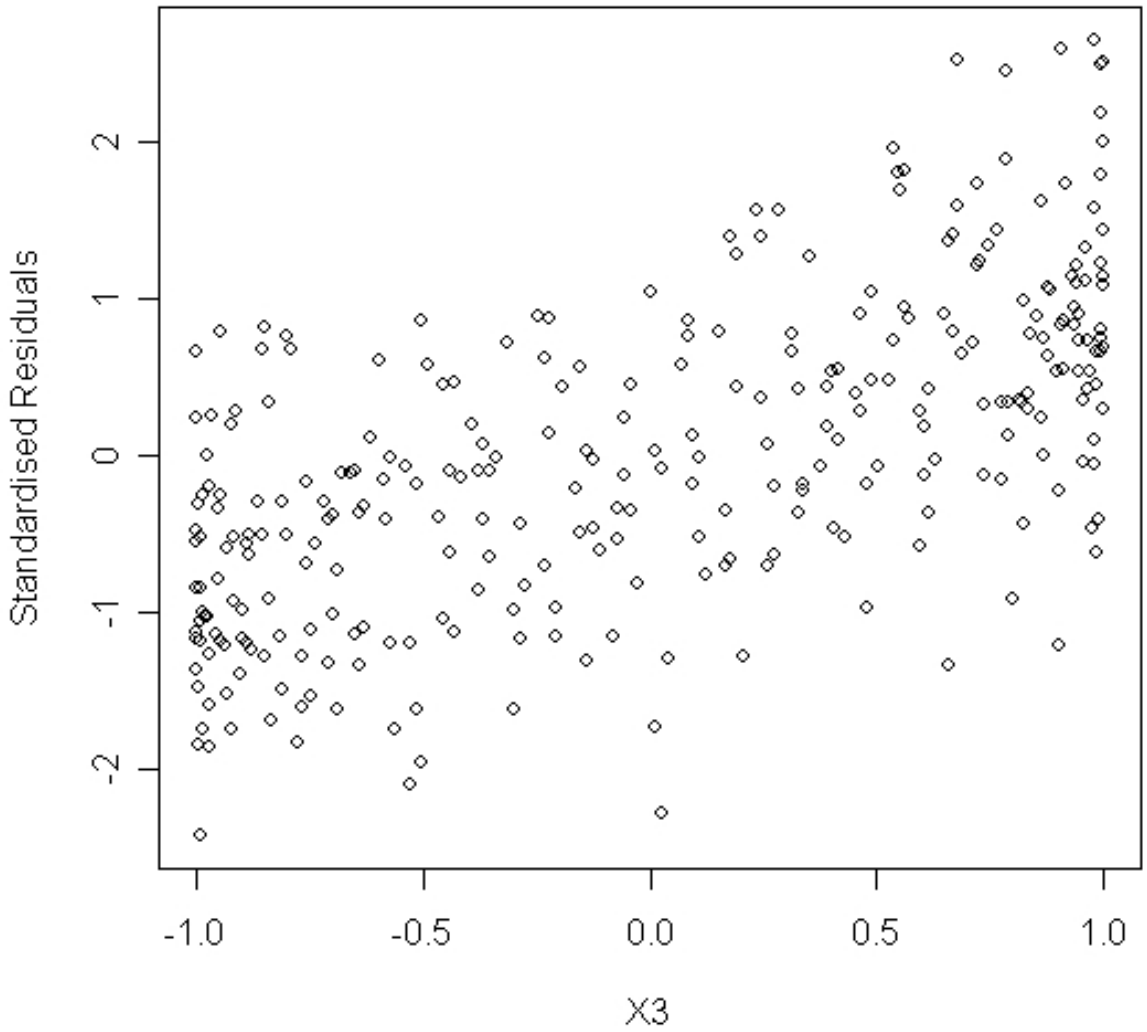




Linearity NOT OK – need to add $\sin(x_1)$ in model



Important Covariate Left out



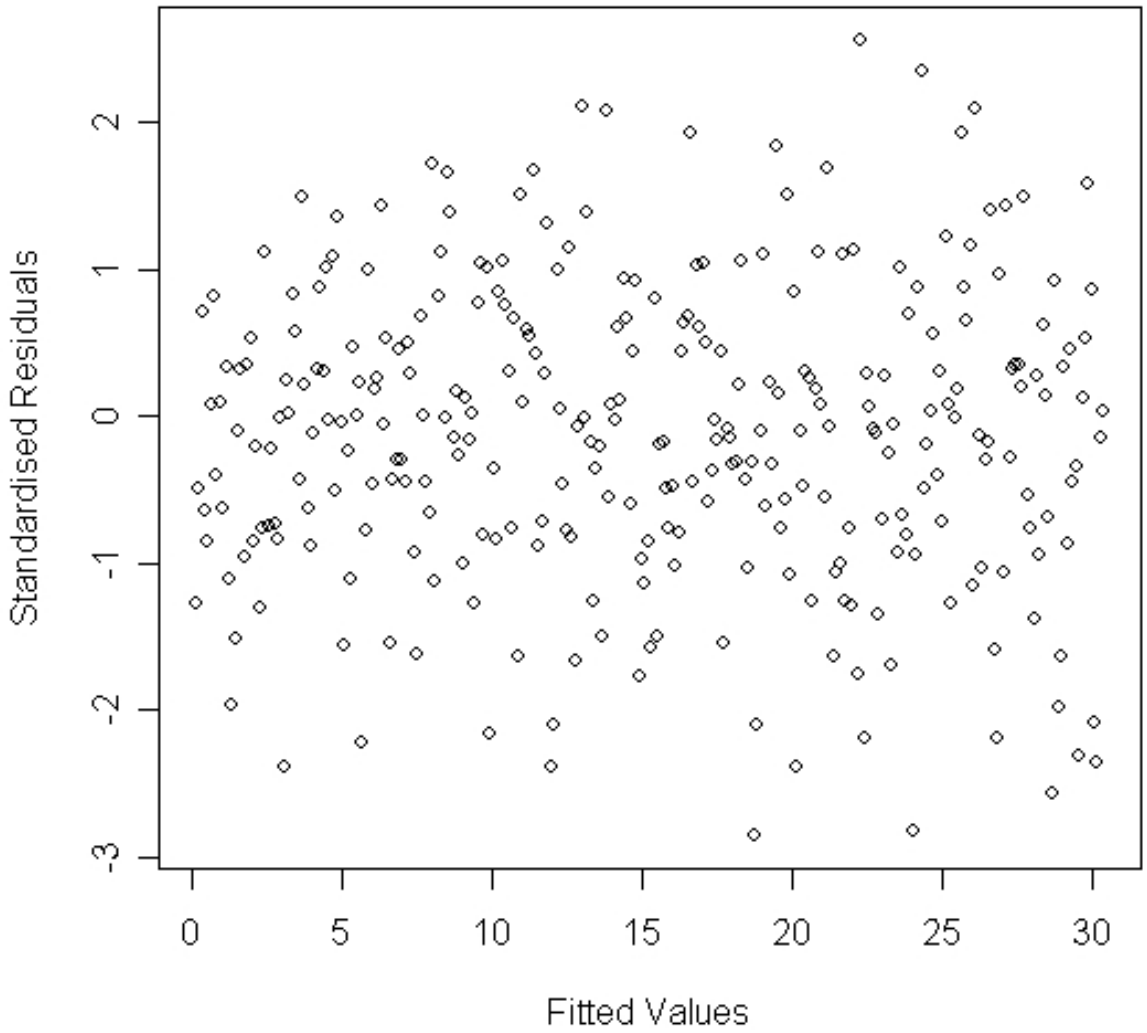
$$\text{Homoskedastic} = \underbrace{\acute{o}\mu o}_{same} + \underbrace{\sigma\kappa\varepsilon\delta\alpha\sigma\mu\grave{o}\varsigma}_{spread}$$

According to our model assumptions, the variance of the errors ε_j should be the same across indices:

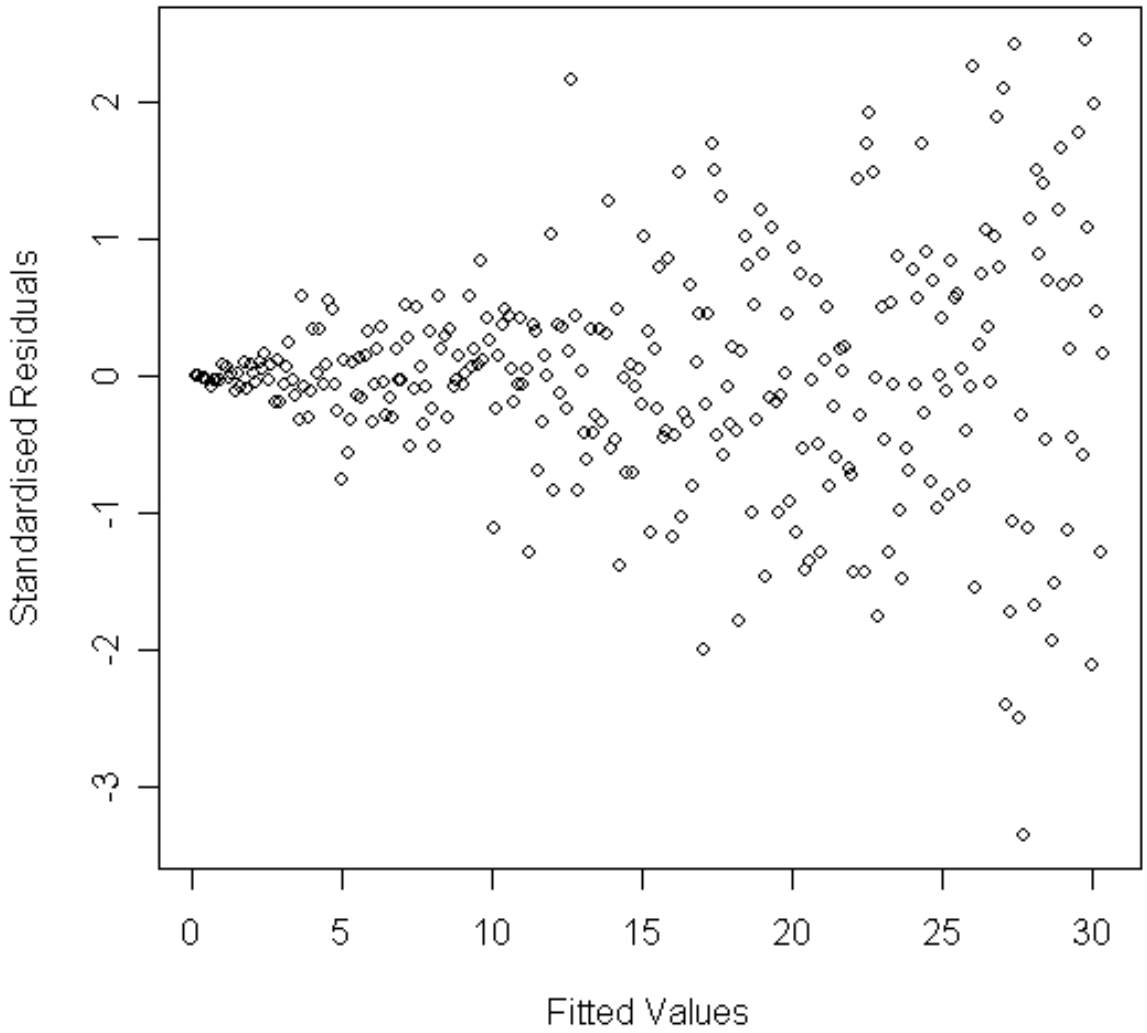
$$\text{var}(\varepsilon_j) = \sigma^2$$

- Plot r against the fitted values $\hat{y}$. (why not against y?)
  - ↪ A random scatter should appear, with approximately constant spread of the values of r for the different values of $\hat{y}$. "Trumpet" or "bulging" effects indicate failure of the homoskedasticity assumption.
  - ↪ Since $\hat{y}^\top e = 0$, this plot can also be used to check linearity, as before.

Homoskedasticity OK



Heteroskedasticity (i.e. lack of Homoskedasticity)



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

How?

Compare the empirical with the theoretical quantiles ...

The *p-quantile* ($p \in [0,1]$) of a distribution F is the value δ defined as

$$\delta := \inf\{\alpha \in \mathbb{R} : F(\alpha) \geq p\}.$$

Notation: $\delta = F^{-1}(p)$ (although the inverse may not be well defined) Given a sample $X_1, \ldots, X_n$, the *empirical p quantile* is the value γ defined as

$$\gamma = \inf\left\{\alpha \in \mathbb{R} : \frac{\#\{X_i \leq \alpha\}}{n} \geq p\right\}.$$

Notation: $\gamma = \hat{F}_n^{-1}(p)$

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

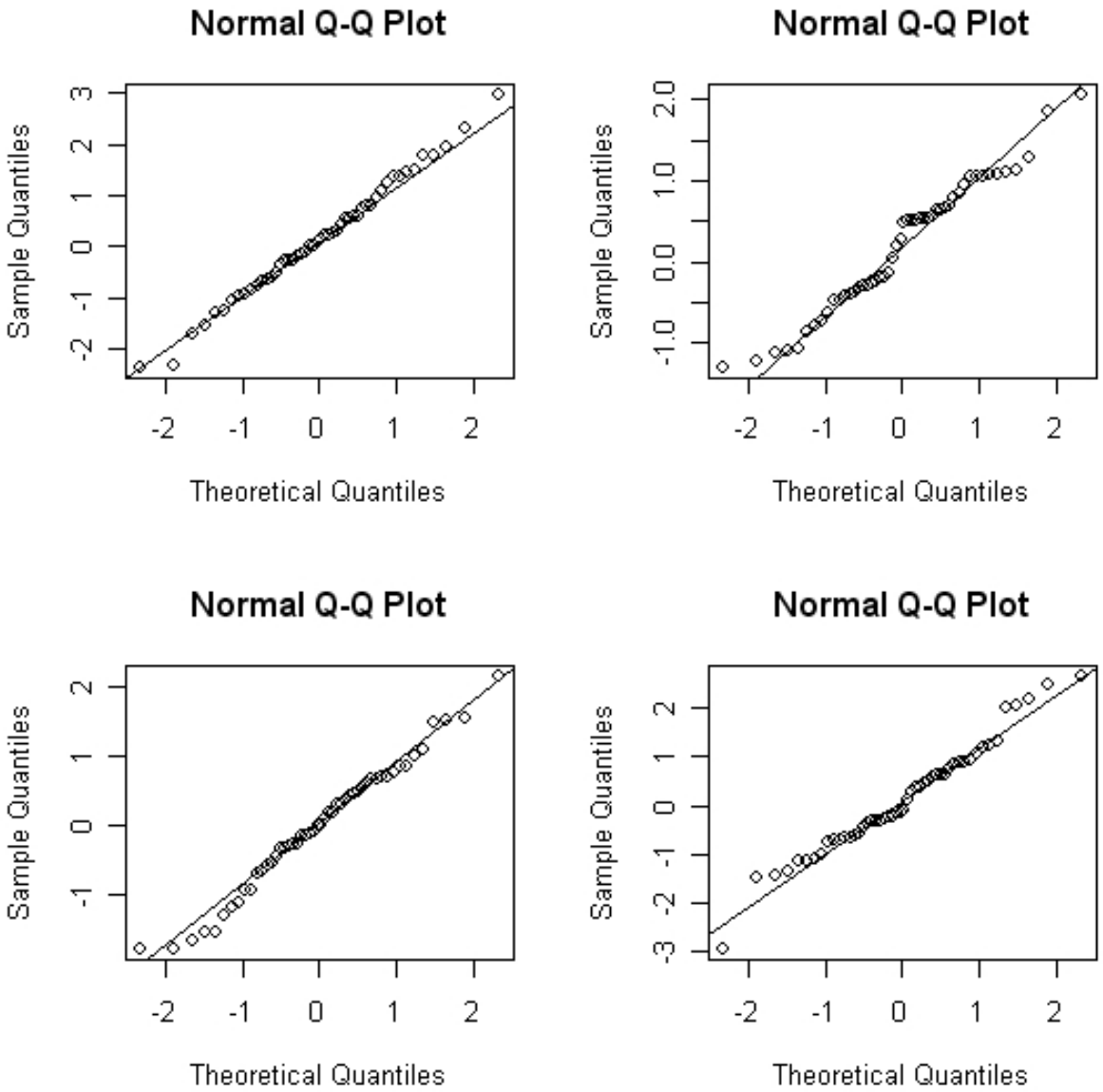
$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$$

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

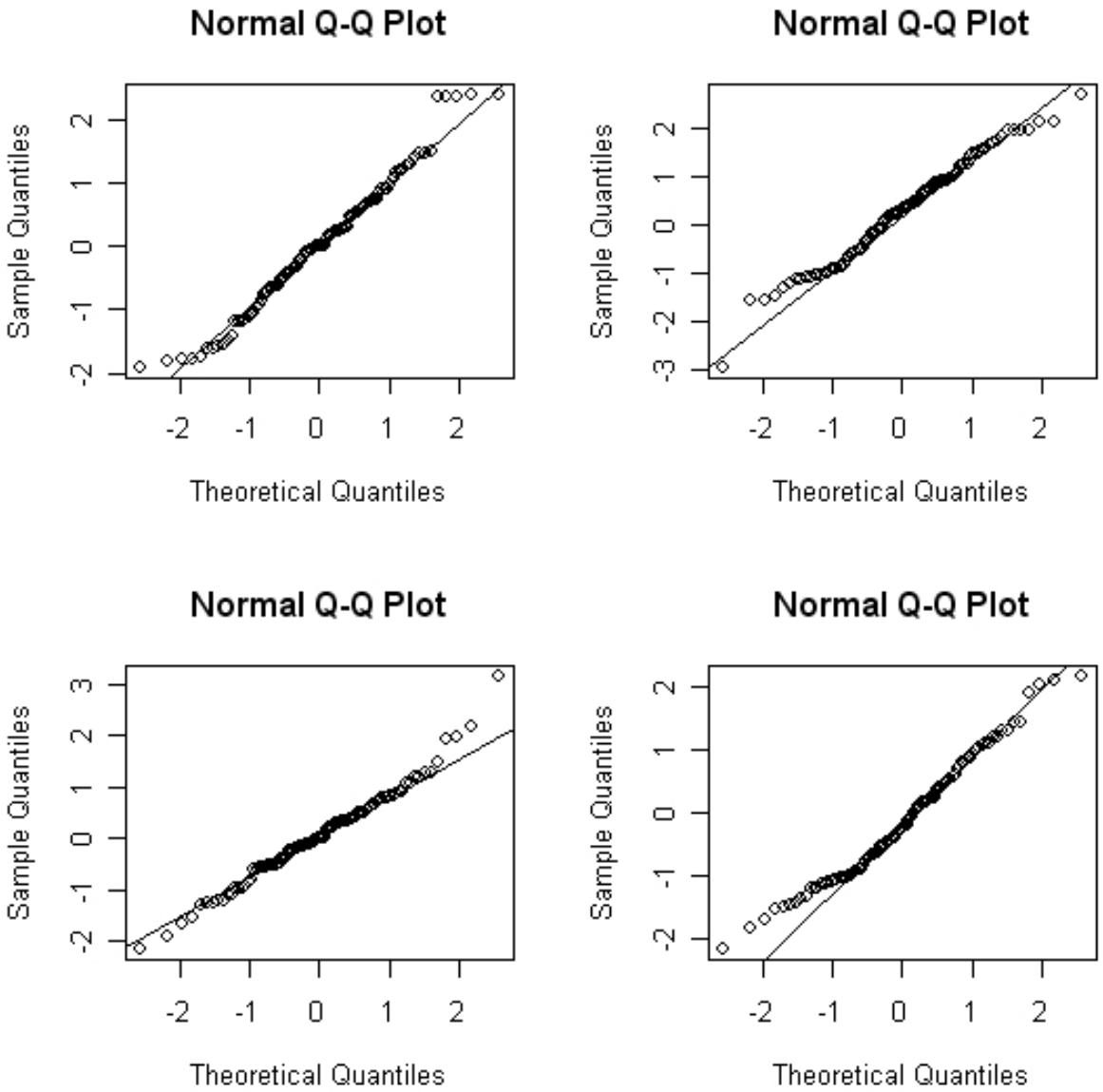
  - ↪ Think why we pick $\Phi^{-1}\left(\frac{k}{n+1}\right)$ instead of $\Phi^{-1}\left(\frac{k}{n}\right)$.
  - ↪ If the points of the quantile plot deviate significantly from the 45° line, there is evidence against the normality assumption. Outliers, skewness and heavy tails easily revealed.

Beware of overinterpretation when n is small!

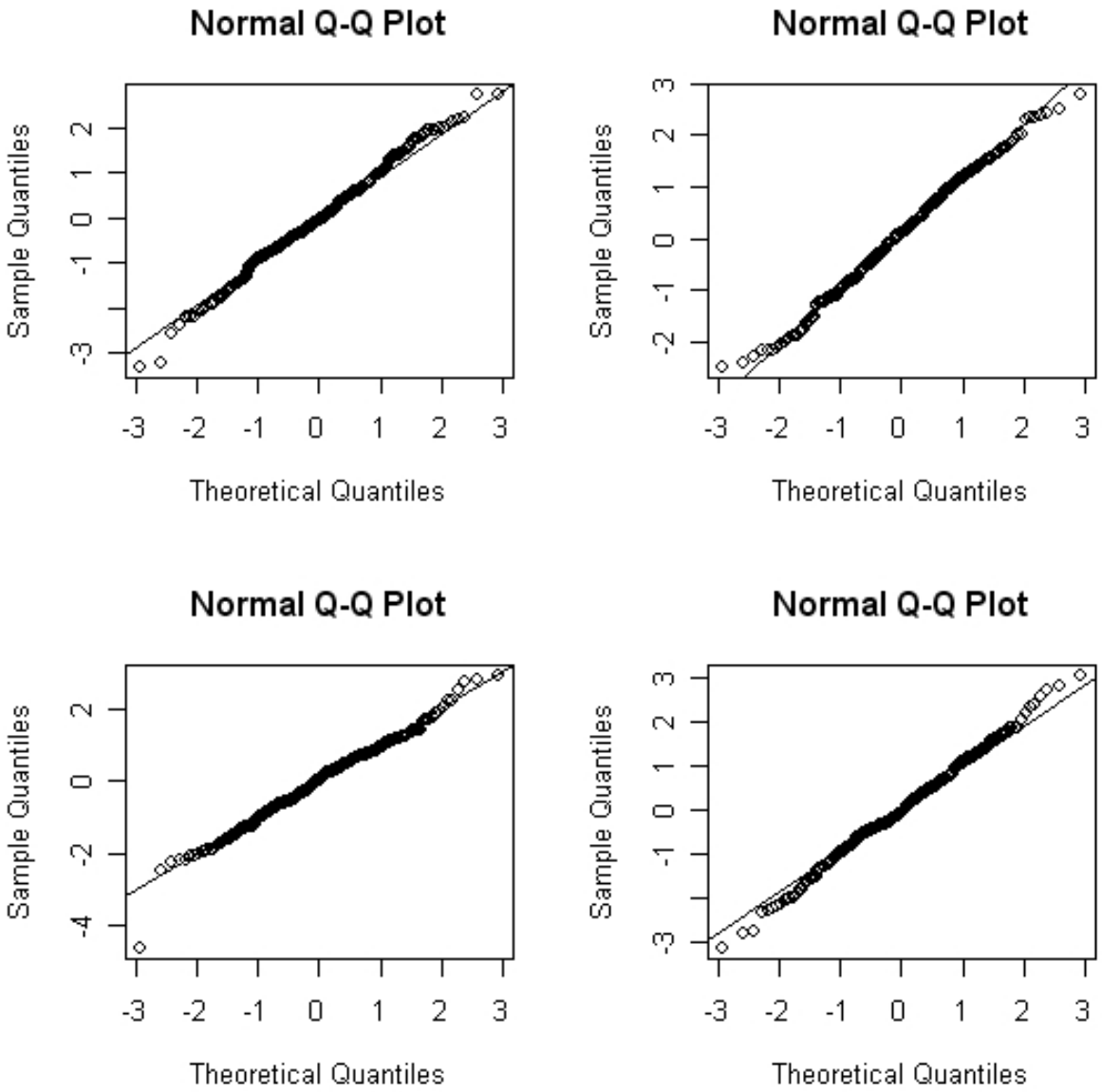
QQ Plot for $n = 50$



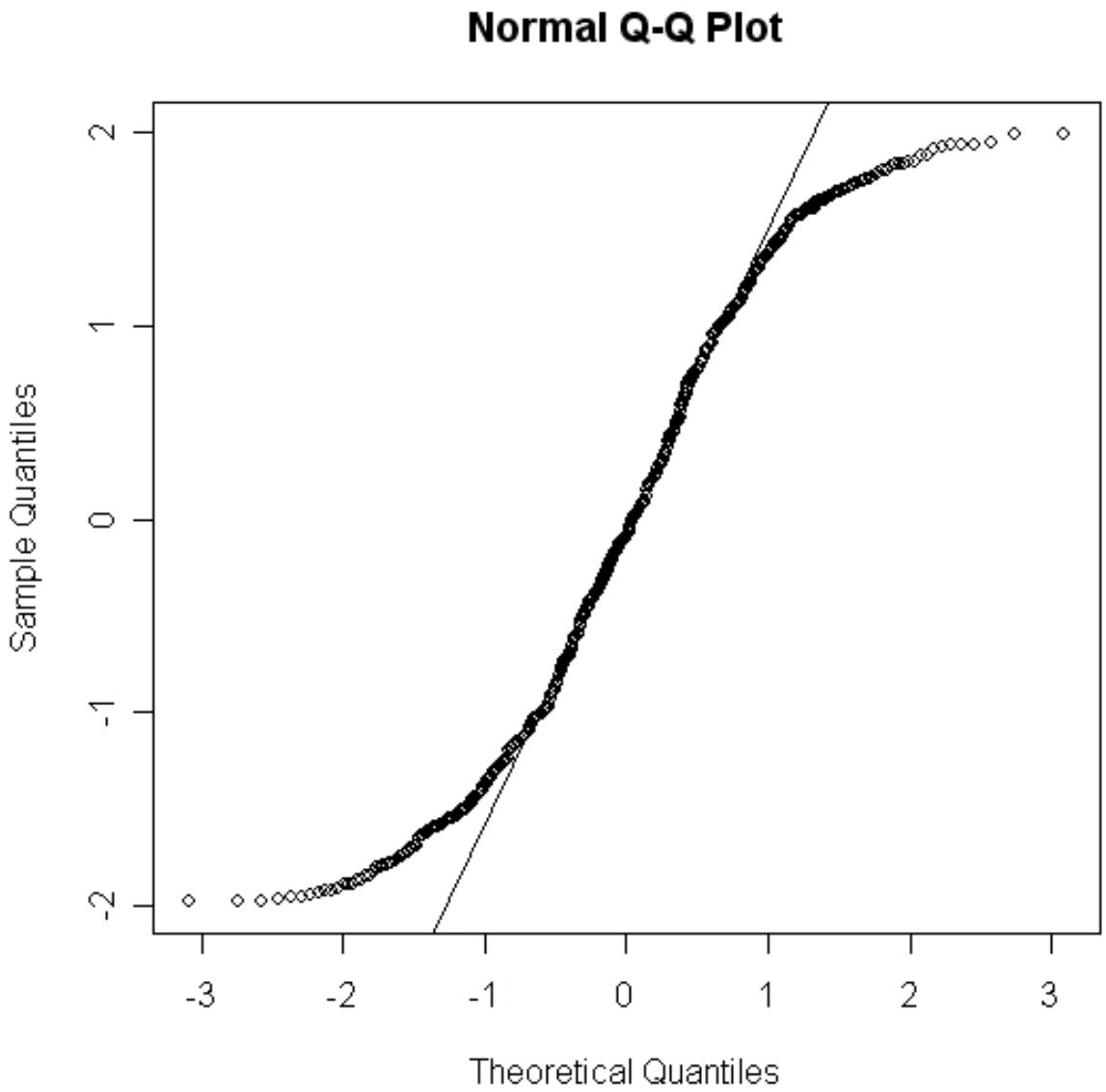
QQ Plot for $n = 100$



QQ Plot for $n = 300$



Normality NOT OK



- It is assumed that $\text{var}[\varepsilon] = \sigma^2 I$.
- Under assumption of normality this is equivalent to independence

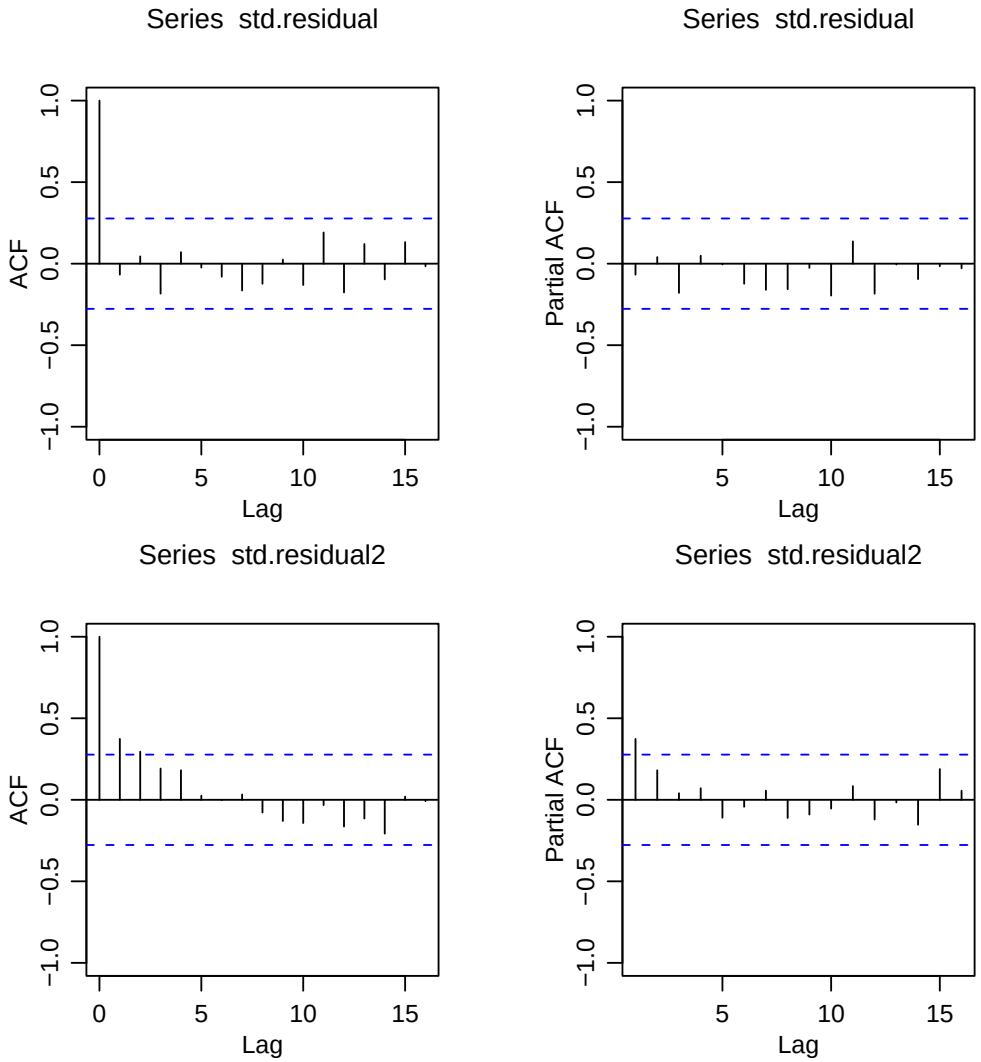
Difficult to check this assumption in practice.

- One thing to check for is clustering, which may suggest dependence.
  - ↪ e.g. identifying groups of related individuals with correlated responses
- When observations are time-ordered can look at correlation $\text{corr}[r_t, r_{t+k}]$ or partial correlation $\text{corr}[r_t, r_{t+k} | r_{t+1}, \ldots, r_{t+k-1}]$. When such correlations exist, we enter the domain of *time series*.

Existence of dependence:

- seriously affects estimator reliability
- inflates standard errors

Checking for Independence



An influential observation can usually be categorised as an:

- *outlier* (relatively easier to spot by eye)

  OR

- *leverage point* (not as easy to spot by eye)

Influential observations

- May or may not decisively affect model validity.
- Require scrutiny on an individual basis and consultation with the data expert.

David Brillinger (Berkeley): *You will not find your Nobel prize in the fit, you will find it in the outliers!*

Influential observations may reveal unanticipated aspects of the scientific problem that are worth studying, and so must not simply be scorned as "non-conformists"!

An *outlier* is an observation that stands out in some way from the rest of the observations, causing *Surprise*! Exact mathematical definition exists (Tukey) but we will not pursue it.

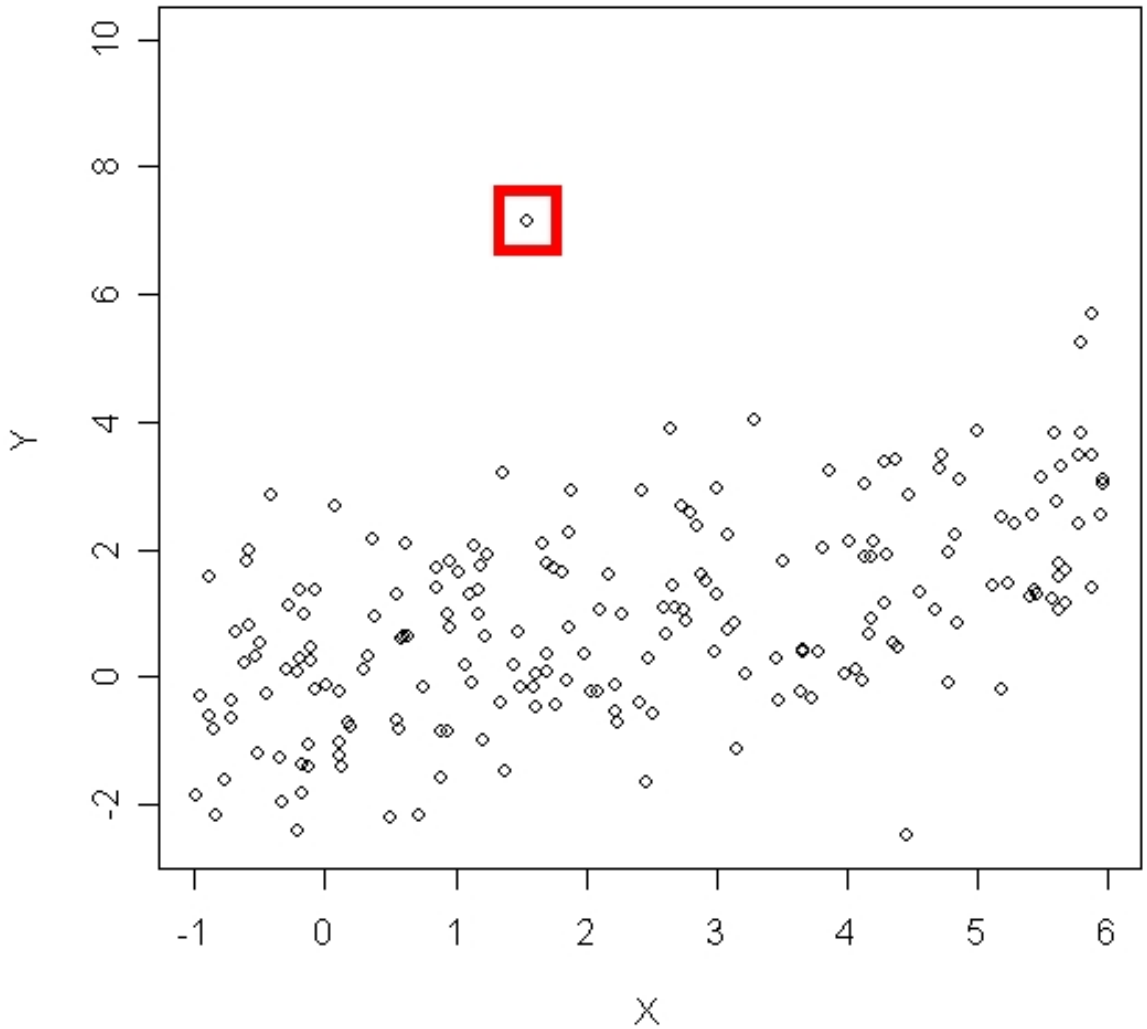
- In regression, outliers are points falling far from the cloud surrounding the regression line (or surface).
- They have the effect of "pulling" the regression line (surface) toward them.

Outliers can be checked for visually through:

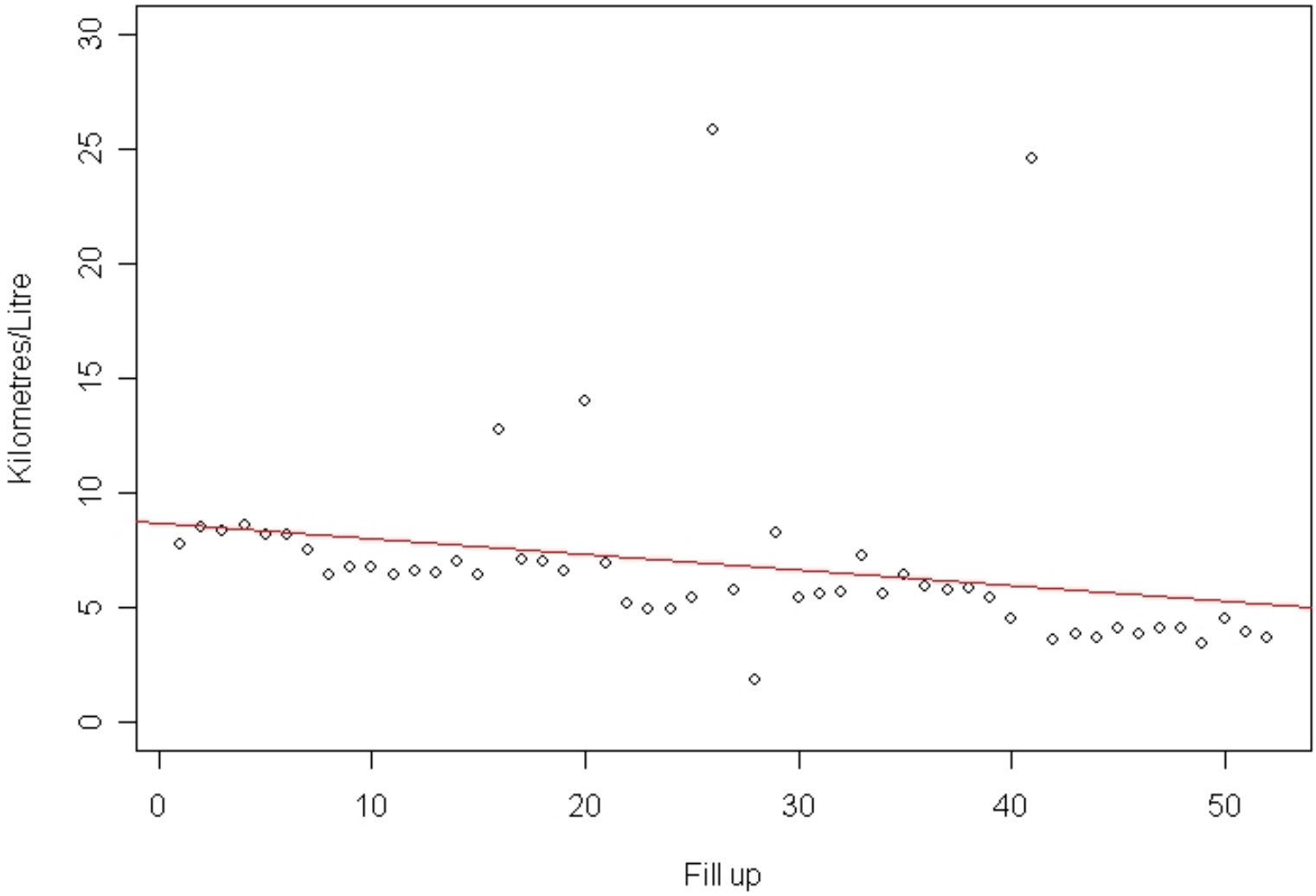
- The regression scatterplot.
  - ↪ Points that can be seen to fall relatively far from the point cloud surrounding the regression line (surface)
- Residual Plots.
  - ↪ Points that fall beyond $(-2, 2)$ in the $(\hat{y}, r)$ plot.

Outliers may result from a data registration error, or a single extreme event. They can, however, result because of a deeper inadequacy of our model (especially if there are many!).

An Outlier



Professor's Van: Outliers



- Outliers may be influential: they "stand out" in the "y-dimension".
- However an observation may also be influential because of unusual values in the "x-dimension".
- Such influential observations cannot be so easily detected through plots.

Call (x_j, y_j) the *j-th case* and notice that

$$\text{var}(y_j - \hat{y}_j) = \text{var}(e_j) = \sigma^2(1 - h_{jj}).$$

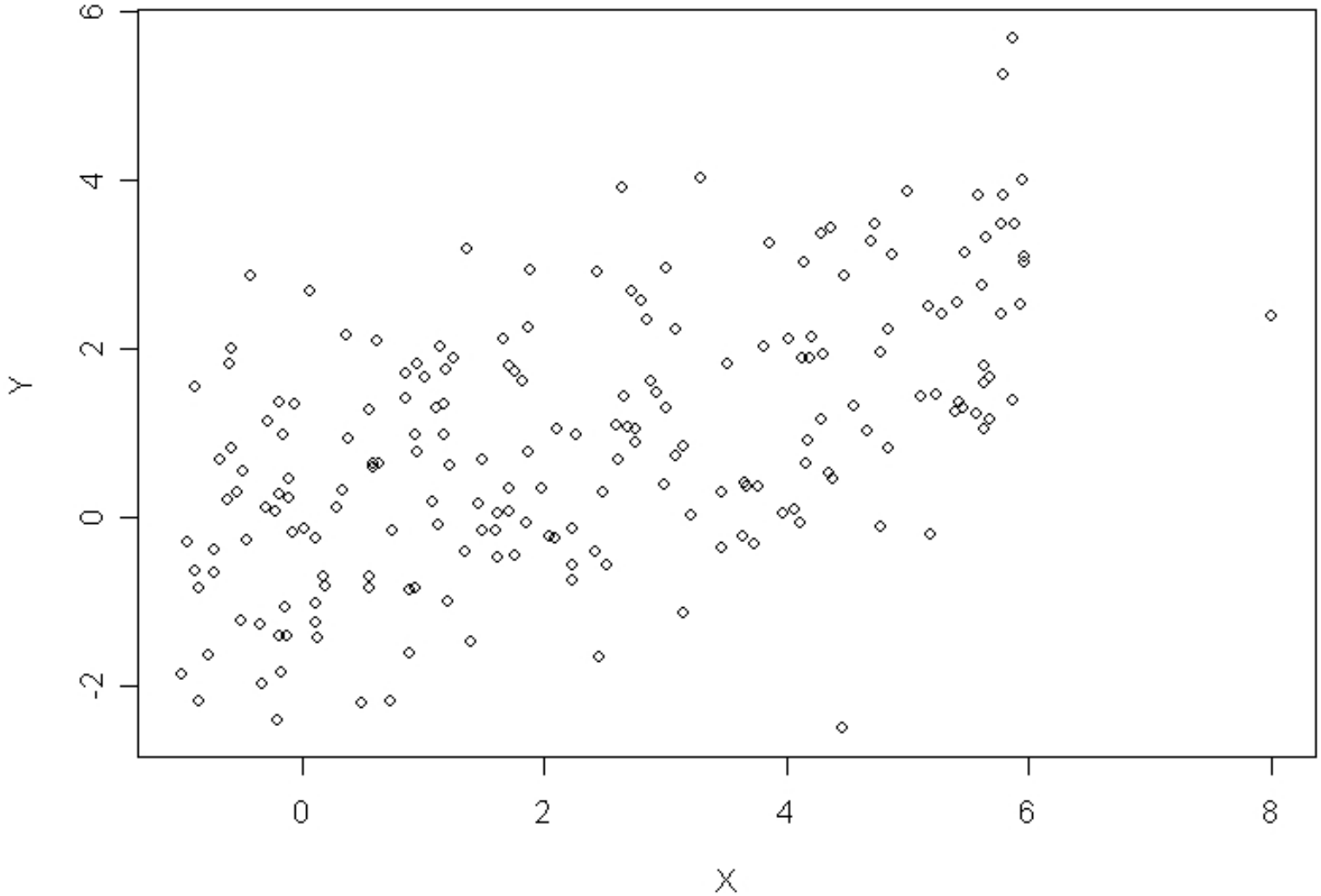
If $h_{jj} \approx 1$, then the model is constrained so $\hat{y}_j = x_j^\top \hat{\beta} \simeq y_j$! (i.e., need a separate parameter entirely devoted to fitting this observation!)

- h_{jj} is called the *leverage* of the j-th case.
- since $\text{trace}(H) = \sum_{j=1}^n h_{jj} = p$, cannot have low leverage for all cases
- a good design corresponds to $h_{jj} \simeq p/n$ for all j

  (i.e. assumption $\max_{j \leq n} h_{jj} \overset{n \to 0}{\to} 0$ satisfied in asymptotic thm).

Leverage point: (rule of thumb) if $h_{jj} > 2p/n$ observation needs further scrutiny—e.g., fitting again without j-th case and studying effect.
Outlier+Leverage Point = TROUBLE

A (very) Noticeable Leverage Point



- How to find cases having strong effect on fitted model?
- Idea: see effect when case j, i.e., (x_j, y_j), is dropped.
- Let $\hat{\beta}_{-j}$ be the LSE when model is fitted to data without case j, and let $\hat{y}_{-j} = X\hat{\beta}_{-j}$ be the corresponding fitted value.
- Define *Cook's distance*

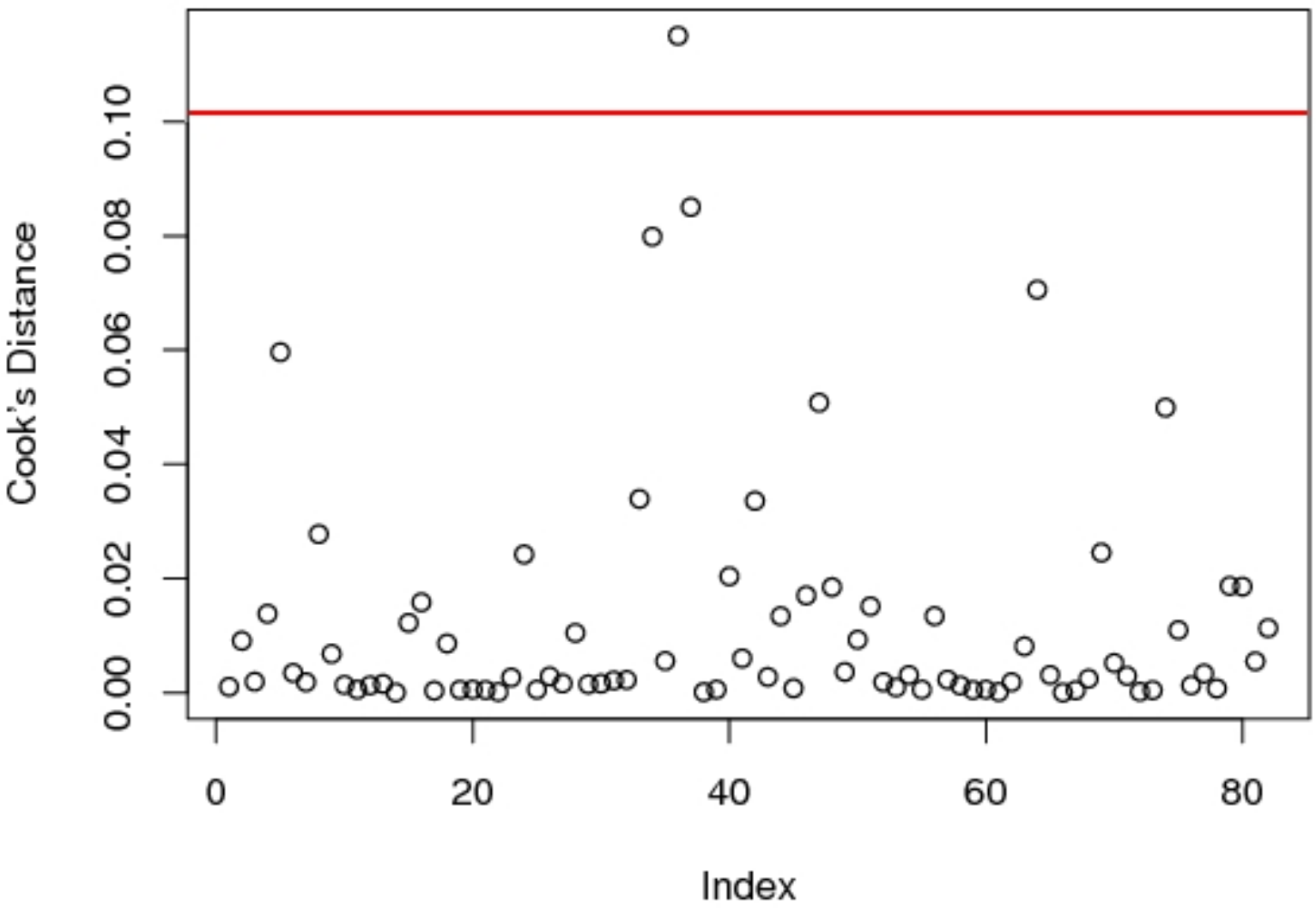
$$C_j = \frac{1}{ps^2}(\hat{y} - \hat{y}_{-j})^\top(\hat{y} - \hat{y}_{-j}),$$

  which measures scaled distance between $\hat{y}$ and $\hat{y}_{-j}$.
- Can show that

$$C_j = \frac{r_j^2 h_{jj}}{p(1 - h_{jj})},$$

  so 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)
- Plot C_j against index $j = 1, \ldots, n$ and compare with $8/(n - 2p)$ level.

A Cook Distance Plot



Diagnostic plots usually constructed:

- y against columns of X
  ↪ check for linearity and outliers
- standardized residual r against columns of X
  ↪ check for linearity
- r against explanatories not included
  ↪ check for variables left out
- r against fitted value $\hat{y}$
  ↪ check for homoskedasticity
- Normal quantile plot
  ↪ check for normality
- Cook's distance plot
  ↪ check for influential observations

Detour: Reminder on Hypothesis Tests

- Scientific theories lead to assertions that are testable using empirical data.
- Data may discredit the theory (call it the *hypothesis*) or not (i.e., empirical findings reasonable under hypothesis).
- Example: Theory of "luminoferous aether" in late 19th century to explain light travelling in vacuum. Discredited by Michelson-Morley experiment.
- Similarities with the logical/mathematical concept of a *necessary condition*.

- H_0: The null hypothesis
  ↪ scientific theory under scrutiny
- $\begin{cases} Y, & \text{data} \\ T(\cdot), & \text{test statistic, assumed positive} \end{cases}$
  ↪ the experimental setup to test theory

INTUITION:

- The null hypothesis would predict a certain plausible range of values for $T(Y)$ (plausible results of the experiment).
- We would say that the assertion made by the null hypothesis (theory) is not supported by the data if $T(Y)$ is an extreme (unlikely) observation given the range of plausible values predicted by the hypothesis (if the experimental evidence appears to be inconsistent with the theory).

Plausibility of different values of $T(\cdot)$ under the theory H_0
$\rightarrow$ described by the distribution of $T(Y)$ under the null hypothesis:

$$\mathbb{P}_{H_0}[T(Y) \in \cdot]$$

Suppose that we perform the experiment $T(Y)$ and the result is $T(Y) = t$. The result t is judged to be incompatible with the hypothesis when

$$p = \mathbb{P}_{H_0}[T(Y) \geq t]$$

is small. The value p is called the *p-value*.

- Small values of p suggest that we have observed something which is unlikely to happen if H_0 holds true.
- Large values of p suggest that what we have observed is plausible if H_0 holds true.
- (Choice of T often guided by an *alternative hypothesis* H_1, under which T should be large)

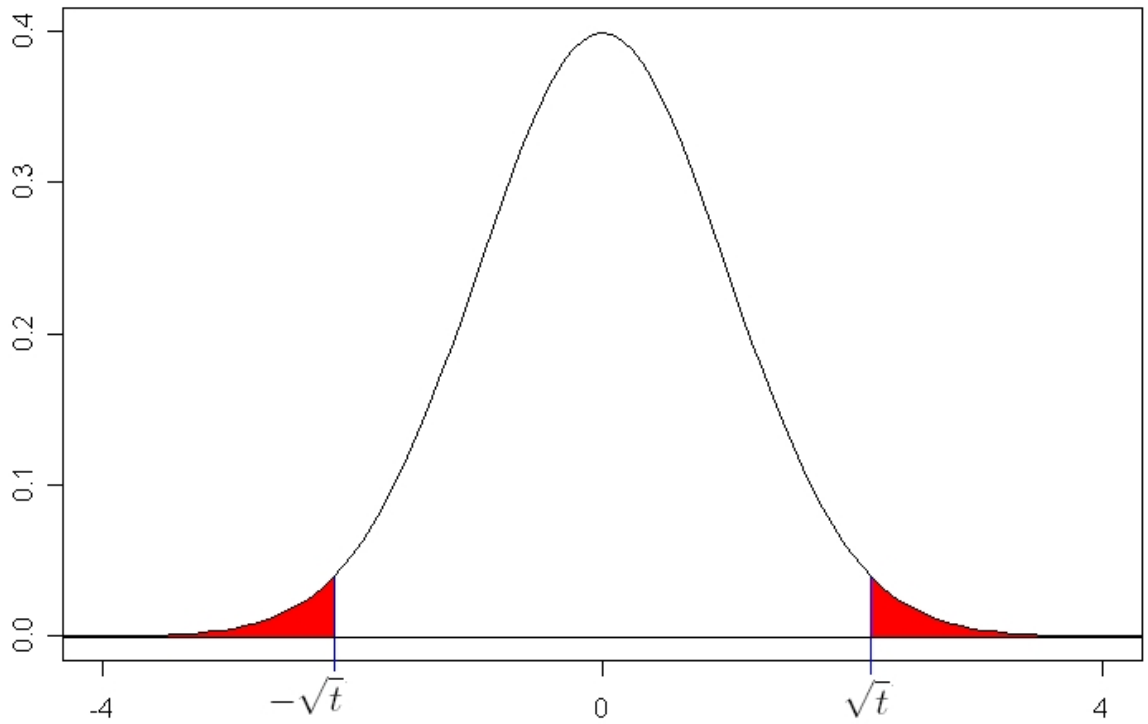
Thus we reject the null hypothesis when p is small.

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$, unknown mean, known variance
- $H_0 : \mu = 0$
- Data: $Y = (X_1, \ldots, X_n)$, $X_i \stackrel{d}{=} X$, X_i indep X_j for $i \neq j$.
- Test statistic: $T(Y) = \left(\frac{\sum_i X_i}{\sigma \sqrt{n}} \right)^2$. (tends to be large when $\mu \neq 0$).
- Perform experiment (i.e., obtain values $y = (x_1, \ldots, x_n)$) and observe $T(y) = t$.

Under the null hypothesis: $T(Y) \stackrel{H_0}{\sim} \chi^2_1$. Hence:

$$\begin{aligned} p &= \mathbb{P}_{H_0}[T(Y) \geq t] \\ &= \mathbb{P}[\chi^2_1 \geq t] \\ &= \mathbb{P}[\{\mathcal{N}(0,1) \leq -\sqrt{t}\} \cup \{\mathcal{N}(0,1) \geq \sqrt{t}\}]. \end{aligned}$$

Usually reject when $p < 0.05$.



- For continuous test statistics with everywhere positive densities, if we reject H_0 whenever $p < \alpha$, then our (type I) error probability is α.
  - ↪ The probability of rejecting H_0 when in fact H_0 is true is α
- There is a close link with confidence intervals.
  - ↪ We will only illustrate this link in a specific example

Example: Testing for $c^\top \beta = 0$ in a Gaussian Regression

- Let $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, unknown β, unknown variance
- $H_0 : c^\top \beta = 0$
- Data: (y, X).
- Test statistic: $T(Y) = \left(\dfrac{c^\top \hat{\beta}}{S\sqrt{c^\top (X^\top X)^{-1} c}} \right)^2$

Suppose we observe $T(y) = \tau$ and let $W \sim t_{n-p}$. Then,

$$p = \mathbb{P}_{H_0}[T(Y) \geq \tau] = \mathbb{P}[\{W \leq -\sqrt{\tau}\} \cup \{W \geq \sqrt{\tau}\}].$$

Reject the null hypothesis if $p < \alpha$, some small α.

- Identical to building a $1 - \alpha$ confidence interval for $c^\top \beta$ based on $\frac{c^\top \hat{\beta} - c^\top \beta}{S\sqrt{c^\top (X^\top X)^{-1} c}}$ and rejecting the hypothesis H_0 if and only if the interval does not contain zero.

Many many issues remain (this was just a reminder!)

- The role of an alternative hypothesis.
- How do we choose a test statistic?
- Are there optimal tests in a given situation?
- Simple and composite hypotheses.
- One and two-sided tests.
- Limitations of hypothesis testing ...
- ...
- Review your 2nd year Probability/Statistics course!

Nested Model Selection & ANOVA

Consider the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

This will always have higher R^2 than the sub-model:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

- Why? (think of geometry...)
- The question is: is the first model *significantly* better than the second one?
  - ↪ i.e. does the first model explain the variation adequately enough, or should we incorporate extra explanatory variables? Need a quantitative answer.

Model is $y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Estimate:

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

Interpretation: $\hat{y} = X\hat{\beta} = Hy$ is the projection of y into the column space of X, $\mathcal{M}(X)$. This subspace has dimension p, when X is of full column rank p.
Now for $q < p$ write X in block notation as

$$X = \underset{n\times q}{X_1} \quad \underset{n\times(p-q)}{X_2}.$$

Interpretation: X_1 is built by the first q columns of X and X_2 by the rest.
Similarly write $\beta = (\beta_1 \ \beta_2)^\top$ so that:

$$y = X\beta + \varepsilon = (X_1 \ X_2)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Our question can now be stated as:

- Is $\beta_2 = 0$?

Let $H_1 = X_1(X_1^\top X_1)^{-1} X_1^\top$, and $\hat{y}_1 = H_1 y$, $e_1 = y - \hat{y}_1$.

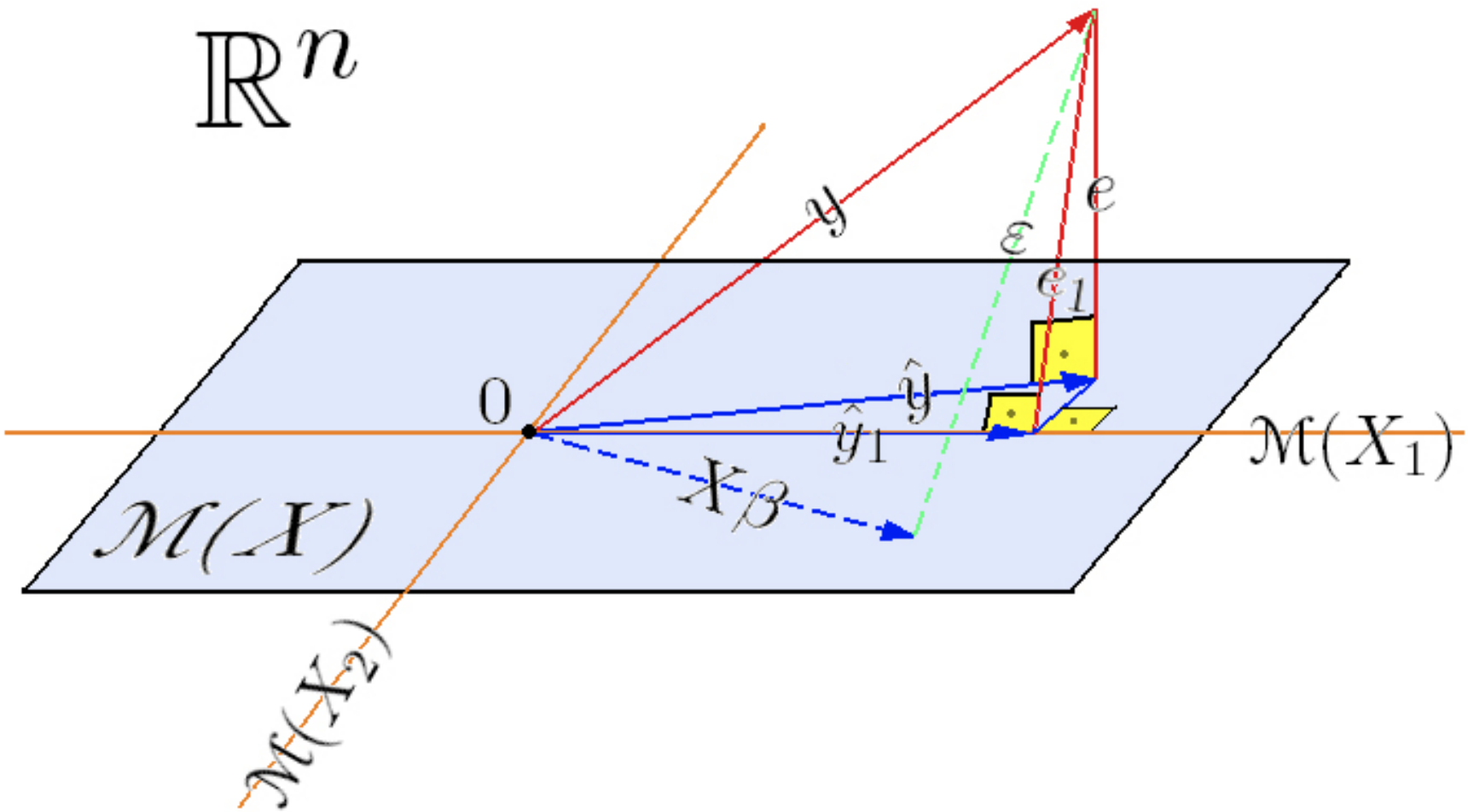
Pythagoras tells us that:

$$\underbrace{\|y - \hat{y}_1\|^2}_{RSS(\hat{\beta}_1) = \|e_1\|^2} = \underbrace{\|y - \hat{y}\|^2}_{RSS(\hat{\beta}) = \|e\|^2} + \underbrace{\|\hat{y} - \hat{y}_1\|^2}_{RSS(\hat{\beta}_1) - RSS(\hat{\beta}) = \|e - e_1\|^2}$$

Notice that $RSS(\hat{\beta}_1) \geq RSS(\hat{\beta})$ always (think why!)

So the idea is simple: to see if it is worthwhile to include β_2 we will compare how much larger $RSS(\hat{\beta}_1)$ is compared to $RSS(\hat{\beta})$.

- Equivalently, we can look at a ratio like $\{RSS(\hat{\beta}_1) - RSS(\hat{\beta})\}/RSS(\hat{\beta})$
- To construct a test based on this quantity, we need to figure out distributions ...


$\mathbb{R}^n$
y
ε
e
e_1
$\hat{y}$
0
$\hat{y}_1$
$\mathcal{M}(X_1)$
$X\beta$
$\mathcal{M}(X)$
$\mathcal{M}(X_2)$

Theorem

We have the following properties:

(A) $e - e_1 \perp e$;

(B) $\|e\|^2 = RSS(\hat{\beta})$ *and* $\|e_1 - e\|^2 = RSS(\hat{\beta}_1) - RSS(\hat{\beta})$ *are independent;*

(C) $\|e\|^2 \sim \sigma^2 \chi^2_{n-p}$;

(D) *under the hypothesis* $H_0 : \beta_2 = 0$, $\|e_1 - e\|^2 \sim \sigma^2 \chi^2_{p-q}$.

Proof.

(A) holds since $e - e_1 = y - \hat{y} - y + \hat{y}_1 = -\hat{y} + \hat{y}_1 \in \mathcal{M}(X_1, X_2)$ but $e \in [\mathcal{M}(X_1, X_2)]^{\perp}$.

To show (B), we notice that

$$e_1 = (I - H_1)y = (I - H_1 H)y$$

because $\mathcal{M}(X_1) \subset \mathcal{M}(X_1, X_2)$.

proof continued

Therefore,

$$e - e_1 = (I - H)y - (I - H_1 H)y = y - Hy - y + H_1 Hy = (H_1 - I)Hy.$$

But recall that $y \sim \mathcal{N}(X\beta, \sigma^2 I)$. Therefore, to prove independence of $e - e_1 = (H_1 - I)Hy$ and $e = (I - H)y$, we need to show that

$$(H_1 - I)H[\sigma^2 I](I - H)^\top = 0.$$

This is clearly the case since $H(I - H) = 0$, proving (B).

(C) follows immediately, since we have already proven last time that $\forall \beta$ (even when $\beta_2 = 0$)

$$RSS(\hat{\beta}) \sim \sigma^2 \chi^2_{n-p}$$

proof continued.

To prove (D), we note that

$$e - e_1 = (H_1 - I)Hy \sim \mathcal{N}\{(H_1 - I)HX\beta, \sigma^2 \underbrace{(H_1 - I)HH^\top(H_1 - I)^\top}_{=H-H_1}\}.$$

But $HX = X(X^\top X)^{-1}X^\top X = X$. So, in block notation,

$$e - e_1 \sim \mathcal{N}((H_1 - I)X_1\beta_1 + (H_1 - I)X_2\beta_2, \sigma^2(H - H_1)).$$

Now $(I - H_1)X_1\beta_1 = 0$ always, since $I - H_1$ projects onto $\mathcal{M}^\perp(X_1)$. Therefore,

$$e - e_1 \sim \mathcal{N}(0, \sigma^2(H - H_1)), \qquad \text{when } \beta_2 = 0.$$

Now observe that $(H - H_1)^\top = (H - H_1)$ and $(H - H_1)^2 = (H - H_1)$ (because $\mathcal{M}(X_1) \subset \mathcal{M}(X_1, X_2)$). Thus,

$$\begin{aligned} e - e_1 \sim \mathcal{N}(0, \sigma^2(H - H_1)^2) \implies e - e_1 &\overset{d}{=} (H - H_1)\varepsilon \\ \implies RSS(\hat{\beta}_1) - RSS(\hat{\beta}) = \|e - e_1\|^2 &\overset{d}{=} \varepsilon^\top(H - H_1)\varepsilon \sim \sigma^2\chi^2_{p-q}. \end{aligned}$$

since $(H - H_1)$ is symmetric idempotent with trace $p - q$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Corollary

We conclude that, under the hypothesis $\beta_2 = 0$,

$$\frac{\left(\frac{RSS(\hat{\beta}_1) - RSS(\hat{\beta})}{p-q}\right)}{\left(\frac{RSS(\hat{\beta})}{n-p}\right)} \sim F_{p-q,n-p}$$

Distributional results suggest the following test:

- Have $Y \sim \mathcal{N}(X_1\beta_1 + X_2\beta_2, \sigma^2 I)$
- $H_0 : \beta_2 = 0$
- Data: (y, X_1, X_2).
- Test statistic: $T = \dfrac{\left(\dfrac{RSS(\hat{\beta}_1) - RSS(\hat{\beta})}{p-q}\right)}{\left(\dfrac{RSS(\hat{\beta})}{n-p}\right)}$

Then, under H_0, it holds that $T \sim F_{p-q,n-p}$. Suppose we observe $T = \tau$. Then,

$$p = \mathbb{P}_{H_0}[T(Y) \geq \tau] = \mathbb{P}[F_{p-q,n-p} \geq \tau]$$

Reject the null hypothesis if $p < \alpha$, some small α, usually 0.05.

- We fitted the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

- But would the following simpler model be in fact adequate?

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Intuitively: is the extra explanatory power of the "larger" model significant enough in order to justify its use instead of a simpler model? (i.e., is the residual vector for the "larger" model significantly smaller than that of the simpler model?)
- In this case, $n = 13$, $p = 5$, $q = 2$ and

$$RSS(\hat{\beta}) = 47.86, \qquad RSS(\hat{\beta}_1) = 1265.7$$

yielding

$$\tau = \frac{(1265.7 - 47.86)/(5-2)}{(47.86)/(13-5)} = 67.86$$

- $p = \mathbb{P}[F_{3,8} \geq 67.86] = 4.95 \times 10^{-6}$, so we reject the hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

- We can fit the quadratic model:

$$\texttt{MPG} = \beta_0 + \beta_1 \texttt{horsepower} + \beta_2 \texttt{horsepower}^2 + \varepsilon$$

- But would the model only with linear term suffice?

$$\texttt{MPG} = \beta_0 + \beta_1 \texttt{horsepower} + \varepsilon$$

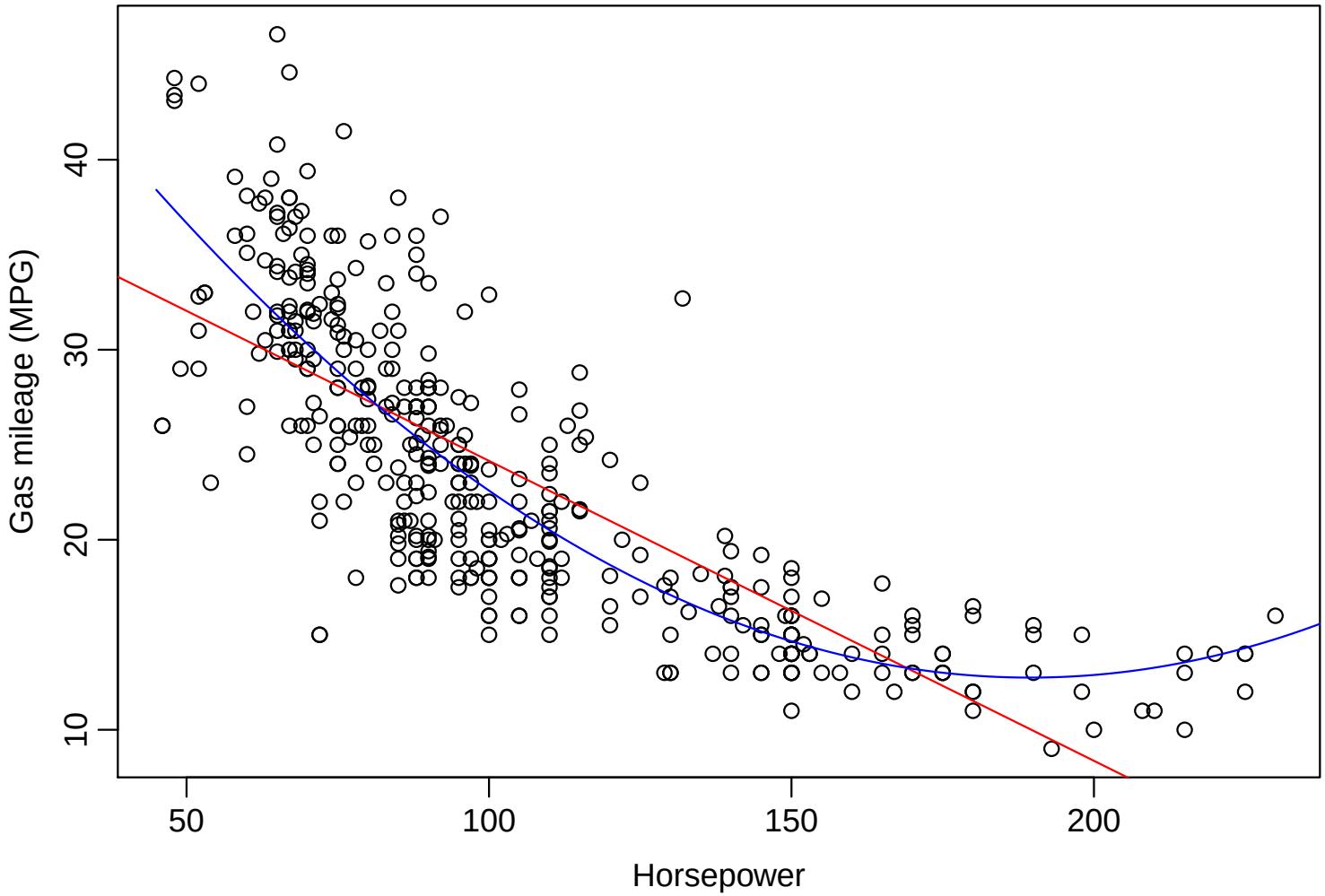
- Intuitively: is the reduction of RSS afforded by the "complex" model substantial enough in order to justify its use instead of a simpler model?
- In this case, $n = 392$, $p = 3$, $q = 2$ and

$$RSS(\hat{\beta}) = 7442, \qquad RSS(\hat{\beta}_1) = 9385.9$$

yielding

$$\tau = \frac{(9385.9 - 7442)/(3-2)}{7442/(392-3)} = 101.6$$

- $p = \mathbb{P}[F_{1,389} \geq 101.6] = 2.2 \times 10^{-21}$, so we reject the hypothesis $H_0 : \beta_2 = 0$.


Gas mileage (MPG)
10
20
30
40
50
100
150
200
Horsepower

- Let $\mathbf{1}$, X_1, ...,X_r be groups of columns of X (the "terms"), such that

$$X = (\underset{n\times 1}{\mathbf{1}} \quad \underset{n\times q_1}{X_1} \quad \underset{n\times q_2}{X_2} \quad \dots \quad \underset{n\times q_r}{X_r}), \quad \beta = (\underset{1\times 1}{\beta_0} \quad \underset{1\times q_1}{\beta_1} \quad \underset{1\times q_2}{\beta_2} \quad \dots \quad \underset{1\times q_r}{\beta_r})^\top$$

We have

$$y = X\beta + \varepsilon = \mathbf{1}\beta_0 + X_1\beta_1 + \cdots + X_r\beta_r + \varepsilon$$

- Would like to do the same "F-test investigation", but this time do it term-by-term. That is, we want to look at the following sequence of nested models:
  - $y = \mathbf{1}\beta_0 + \varepsilon$
  - $y = \mathbf{1}\beta_0 + X_1\beta_1 + \varepsilon$
  - $y = \mathbf{1}\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$
  - $\vdots$
  - $y = \mathbf{1}\beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_r\beta_r + \varepsilon$

Proceed similarly as before. Define:

- $X_0 := \mathbf{1}$ and $\mathcal{X}_k = (X_0 \; X_1 \; X_2 \; \dots \; X_k), \quad k \in \{0, \dots, r\}$
- $\mathcal{H}_k := \mathcal{X}_k (\mathcal{X}_k^\top \mathcal{X}_k)^{-1} \mathcal{X}_k^\top, \quad k \in \{0, \dots, r\}$
- $\hat{y}_k := \mathcal{H}_k y, \quad k \in \{0, \dots, r\}$
- $e_k = y - \hat{y}_k, \quad k \in \{0, \dots, r\}$
- Note that $\hat{y}_0 = \bar{y}\mathbf{1}$.

▶ As before, Pythagoras implies

$$\underbrace{\|y - \hat{y}_0\|^2}_{\|e_0\|^2} = \underbrace{\|y - \hat{y}_r\|^2}_{\|e_r\|^2} + \underbrace{\underbrace{\|\hat{y} - \hat{y}_{r-1}\|^2}_{\|e_r - e_{r-1}\|^2} + \cdots + \underbrace{\|\hat{y}_1 - \hat{y}_0\|^2}_{\|e_1 - e_0\|^2}}_{\|e_r - e_0\|^2}$$

$$= \underbrace{\|e_r\|^2}_{RSS_r} + \sum_{k=0}^{r-1} \underbrace{\|e_{k+1} - e_k\|^2}_{RSS_k - RSS_{k+1}}$$

with RSS_k the residual sum of squares for $\hat{y}_k$, with ν_k degrees of freedom.

Some observations:

- $RSS_k - RSS_{k+1}$ is the reduction in residual sum of squares caused by adding X_{k+1}, when the model already contains $X_0, \ldots, X_k$.
- RSS_r and $\{RSS_k - RSS_{k+1}\}_{k=0}^{r-1}$ are all mutually independent.
- Obviously, $\nu_0 \geq \nu_1 \geq \nu_2 \geq \cdots \geq \nu_r$
- $\nu_{k+1} = \nu_k$ if $X_{k+1} \in \mathcal{M}(\mathcal{X}_k)$.

▶Given this information, we want to see how adding each term in the model sequentially, affects the explanatory capacity of the model.

↪ In other words, we want to investigate the reduction in the residual sum of squares achieved by adding each term to the model. Is this significant?

Terms	df	Residual RSS	Terms added	df	Reduction in RSS	F-test
$\mathbf{1}$	$n-1$	RSS_0				
$\mathbf{1}, X_1$	ν_1	RSS_1	X_1	$n-1-\nu_1$	$RSS_0 - RSS_1$	
$\mathbf{1}, X_1, X_2$	ν_2	RSS_2	X_2	$\nu_1 - \nu_2$	$RSS_1 - RSS_2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\mathbf{1}, X_1, \ldots, X_r$	ν_r	RSS_r	X_r	$\nu_{r-1} - \nu_r$	$RSS_{r-1} - RSS_r$	

The F-statistic for testing the significance of the reduction in RSS when X_k is added to the model containing terms $1, X_1, \ldots, X_k$ is

$$F_k = \frac{(RSS_{k-1} - RSS_k)/(\nu_{k-1} - \nu_k)}{RSS_r/\nu_r},$$

and $F_k \sim F_{\nu_{k-1}-\nu_k, \nu_r}$ under the null hypothesis $H_0 : \beta_k = 0$.

Large values of F_k relative to the null distribution are evidence against H_0.

- Reductions in overall sum of squares when sequentially entering terms x_1, x_2, x_3 and x_4.
- Does adding extra variables improve model significantly?

	Df	Red Sum Sq	F value (τ)	p-value
x_1	1	1450.08	242.37	2.88×10^{-7}
x_2	1	1207.78	201.87	5.86×10^{-7}
x_3	1	9.79	1.64	0.2366
x_4	1	0.25	0.04	0.8441
Residual SSq	8	47.86		

- In this case, each term is a single column (variable).

Warning!

- Significance of entering a term depends on how the sequence is defined: when entering terms in different order get different results! (why?)
- When a term is entered "early" and is significant, this does not tell us much (why?)
- When a term is entered "late" is significant, then this is quite informative (why?)

▶ Why is this true? Are there special cases when the order of entering terms doesn't matter?

- Consider terms $X_0 = \mathbf{1}, X_1, X_2$ from X, so

$$X = \underset{n\times 1}{(\,X_0} \quad \underset{n\times q_1}{X_1} \quad \underset{n\times q_2}{X_2\,)}, \quad \beta = \underset{1\times 1}{(\,\beta_0} \quad \underset{1\times q_1}{\beta_1} \quad \underset{1\times q_2}{\beta_2\,)^\top}$$

- Assume orthogonality of terms, i.e. $X_i^\top X_j = 0, \quad i \neq j$

Notice that in this case

$$\hat{\beta} = \begin{pmatrix} X_0^\top X_0 & 0 & 0 \\ 0 & X_1^\top X_1 & 0 \\ 0 & 0 & X_2^\top X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_0 & X_1 & X_2 \end{pmatrix}^\top y$$

$$\implies \hat{\beta}_0 = \bar{y},\ \hat{\beta}_1 = (X_1^\top X_1)^{-1} X_1^\top y,\ \hat{\beta}_2 = (X_2^\top X_2)^{-1} X_2^\top y$$

It follows that the reductions of sums of squares are unique, in the sense that they do not depend upon the order of entry of the terms in the model. (show this!)
Intuition: X_i contains completely independent linear information from X_j for y, $i \neq j$

Model Selection / Collinearity / Shrinkage

- **Theory:** We are given a relationship

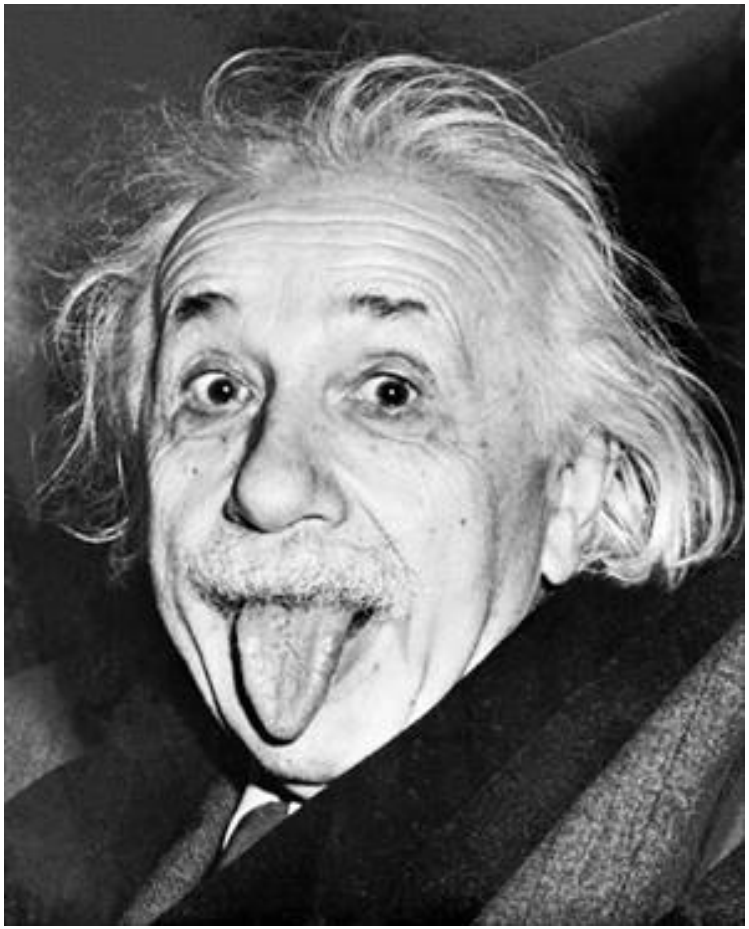
$$y = X\beta + \varepsilon$$

and asked to provide estimators, tests, confidence intervals, optimality properties ...

... and we can do it with complete success!

- **Practice**: We are given data (y, X) and suspect a linear relationship between y and some of the columns of X. We don't know a priori which exactly!
  - ↪ Need to select a "most appropriate" subset of the columns of X
  - ↪ General principle: parsimony (Latin *parsimōnia*: sparingness; simplicity and least number of requisites and assumptions; economy or frugality of components and associations).

Albert Einstein (1879–1955)



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

William of Ockham (?1285–1347)

Occam's razor: *It is vain to do with more what can be done with fewer*.
Given several explanations of the same phenomenon, we should prefer the simplest.

Graphical exploration $\rightsquigarrow$ provides initial picture:

- plots of y against candidate variables;
- plots of transformations of y against candidate variables;
- plots of transformations of certain variables against y;
- plots of pairs of candidate variables.

This will often provide a starting point, but:

- **Automatic Model Selection**: Need objective model comparison criteria, as a screening device.
  - $\hookrightarrow$ We saw how to do an F-test, but what if models to be compared are not nested?
- **Automatic Model Building**: Situations when p large, so there are *lots* of possible models.
  - $\hookrightarrow$ Automatic methods for building a model? We saw that ANOVA depends on the order of entry of variables in the model ...

Consider design matrix X with p variables.

- 2^p possible models!
- Denote set of all models generated by X by 2^X (model powerset)
- If wish to consider k different transformations of each variable, then p becomes $(1+k)p$
- Fast algorithms (branch and bound, `leaps` in `R`) exist to fit them, but they don't work for *large* p, and anyway ...
- ... need criterion for comparison.

So given a collection of models, we need an automatic (objective) way to pick out a "best" one (unfortunately cannot look carefully at all of them, BUT NOTHING replaces careful scrutiny of the final model by an experienced researcher).

Many possible choices, none universally accepted. Some (classical) possibilities:

- Prediction error based criteria (CV)
- Information criteria (AIC, BIC, ...)
- Mallow's C_p statistic

Before looking at these, let's introduce terminology: Suppose that the truth is

$y = X\beta + \varepsilon$ but with $\beta_r = 0$ for some subset β_r of β.

- The true model contains only the columns for which $\beta_r \neq 0$
  - ↪ Equivalently, the true model uses $X_\heartsuit$ as the design matrix, the latter being the matrix of columns of X corresponding to non-zero coefficients.
- A correct model is the true model plus extra columns.
  - ↪ Equivalently, a correct model has a design matrix $X_\diamondsuit$, such that $\mathcal{M}(X_\heartsuit) \subset \mathcal{M}(X_\diamondsuit)$.
- A wrong model is a model that does not contain all the columns of the true model.
  - ↪ Equivalently, a wrong model has a design matrix $X_\diamondsuit$, such that $\mathcal{M}(X_\heartsuit) \cap \mathcal{M}(X_\diamondsuit) \neq \mathcal{M}(X_\heartsuit)$.

▶ We may wish to choose a model by minimising the error we make on average, when predicting a future observation given our model.
Our "experiment is":

- Design matrix X
- response y at X

Every model $f \in 2^X$, will yield fitted values $\hat{y}(f) = H_f y$. And suppose we now obtain new independent responses y_+ for the same "experimental setup" X.
Then, one approach is to select the model

$$f^* = \arg\min_{f \in 2^X} \underbrace{\frac{1}{n} \mathbb{E}\left\{\|y_+ - \hat{y}(f)\|^2\right\}}_{\Delta(f)},$$

where expectation is taken over both y and y_+.

Let X be a design matrix, and let $X_\diamondsuit$ $(n \times p)$ and $X_\heartsuit$ $(n \times q)$ be matrices built using columns of X. Suppose that the true relationship between y and X is

$$y = \underbrace{X_\heartsuit \beta}_{\mu} + \varepsilon$$

but we use the matrix $X_\diamondsuit$ instead of $X_\heartsuit$ (i.e., we fit a different model). Therefore our fitted values are

$$\hat{y} = (X_\diamondsuit^\top X_\diamondsuit)^{-1} X_\diamondsuit^\top y = H_\diamondsuit y.$$

Now suppose that we obtain new observations y_+ corresponding to the same design X

$$y_+ = X_\heartsuit \beta + \varepsilon_+ = \mu + \varepsilon_+.$$

Then, observe that

$$\begin{aligned} y_+ - \hat{y} &= \mu + \varepsilon_+ - H_\diamondsuit(\mu + \varepsilon) \\ &= (I - H_\diamondsuit)\mu + \varepsilon_+ - H_\diamondsuit \varepsilon. \end{aligned}$$

It follows that

$$\begin{aligned} \|y_+ - \hat{y}\|^2 &= (y_+ - \hat{y})^\top (y_+ - \hat{y}) \\ &= \mu^\top (I - H_\diamondsuit)\mu + \varepsilon^\top H_\diamondsuit \varepsilon + \varepsilon_+^\top \varepsilon_+ + [\text{cross terms}]. \end{aligned}$$

Since $\mathbb{E}[\text{cross terms}] = 0$ (why?), we observe that

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

- Selecting a *correct model* instead of the *true model* brings in additional variance, because $q < p$.
- Selecting a *wrong model* instead of the *true model* results in bias, since $(I - H_\diamondsuit)\mu \neq 0$ when μ is not in the column space of $X_\diamondsuit$.
- Must find a balance between small variance (few columns in the model) and small bias (all columns in the model).

▶ Impossible to calculate Δ (depends on unknown μ and σ^2), so we must find a proxy (estimator) $\widehat{\Delta}$.
Suppose that n is large so that we can split the data in two pieces:

- X^*, y^* used to estimate the model
- X', y' used to estimate the prediction error for the model

The estimator of the prediction error will be

$$\widehat{\Delta} = (n')^{-1}\|y' - X'\hat{\beta}^*\|^2.$$

In practice n can be small and we often cannot afford to split the data (variance of $\hat{\Delta}$ is too large).
Instead we use the *leave-one-out cross validation* sum of squares:

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

where $\hat{\beta}_{-j}$ is the estimate produced when dropping the jth case.

No need to perform n regressions since

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

so the full regression may be used (show this!). Alternatively one may use a more stable version:

$$GCV = \sum_{j=1}^{n} \frac{(y_j - x_j^\top \hat{\beta})^2}{(1 - \text{trace}(H)/n)^2},$$

where "G" stands for "generalised", and we guard against any $h_{jj} \approx 1$.
It holds that:

$$\mathbb{E}[GCV] = \frac{\mu^\top (I - H)\mu}{(1 - p/n)^2} + \frac{n\sigma^2}{1 - p/n} \approx n\Delta.$$

▷ Suggests strategy: pick variables to minimise (G)CV.

Criteria can be obtained based on the notion of *information (relative entropy)*.

- Same basic idea as for prediction error: aim to choose candidate model $f(y)$ to minimise *information distance*:

$$\int \log \left\{ \frac{g(y)}{f(y)} \right\} g(y)\, dy \geq 0,$$

where $g(y)$ represents true model—equivalent to maximising expected log likelihood

$$\int \log f(y) g(y)\, dy.$$

- Can show that (apart from constants) information distance is estimated by

$$\text{AIC} = -2\hat{\ell} + 2p \quad (\equiv n \log \hat{\sigma}^2 + 2p \text{ in linear model})$$

where $\hat{\ell}$ is maximised log likelihood for given model, and p is number of parameters.

- Improved (corrected) version of AIC for regression problems:

$$\text{AIC}_c \equiv \text{AIC} + \frac{2p(p+1)}{n-p-1}.$$

- Also can use *Bayes' information criterion*

$$\text{BIC} = -2\hat{\ell} + p \log n.$$

- Mallows suggested

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

  where SS_p is RSS for fitted model and s^2 estimates σ^2.
- Comments:
  - AIC tends to choose models that are too complicated, buts AIC_c cures this somewhat;
  - BIC is *model selection consistent*—if the true model is among those fitted, BIC chooses it with probability $\to 1$ as $n \to \infty$ (for fixed p).

For each $n \in \{10, 20, 40\}$ we construct 20 $n \times 7$ design matrices. We multiply each of these design matrices from the right with $\beta = (1, 2, 3, 0, 0, 0, 0)^\top$ and we add a $n \times 1$ Gaussian error. We do this independently 50 times, obtaining 1000 regressions with $p = 3$. Selected models with 1 or 2 covariates have a bias term, and those with 4 or more covariates have excess variance.

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
	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
	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
	AIC_c			786	105	52	41	16

- We saw so far:

Automatic Model Selection: build a set of models and select the "best" one.

- Now look at different philosophy:

Automatic Model Building: construct a single model in a way that would hopefully provide a good one.

There are three standard methods for doing this:

- Forward Selection
- Backward Elimination
- Stepwise Selection

CAUTION: Although widely used, these have no theoretical basis. Element of arbitrariness ...

- *Forward selection*: starting from the model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is significant, stop; otherwise
  3. update the current model to include the most significant new term; go to step 1.

- *Backward elimination*: starting from the model with all terms,
  1. if all terms are significant, stop; otherwise
  2. update current model by dropping the term with the smallest F statistic; go to step 1.

- *Stepwise*: starting from an arbitary model,
  1. consider three options—add a term, delete a term, swap a term in the model for one not in the model, and choose the most significant option;
  2. if model unchanged, stop; otherwise go to step 1.

Some thoughts:

- Each procedure may produce a different model.
- Systematic search minimising Prediction Error, AIC or similar over all possible models is preferable— BUT not always feasible (e.g., when p large).
- Stepwise methods can fit 'highly significant' models to purely random data! Main problem is lack of objective function.
- Can be improved by comparing Prediction Error/AIC for different models at each step — uses objective function, but no systematic search.

Data on light water reactors (LWR) constructed in the USA. The covariates are `date` (date construction permit issued), `T1` (time between application for and issue of permit), `T2` (time between issue of operating license and construction permit), `capacity` (power plant capacity in MWe), `PR` (=1 if LWR already present on site), `NE` (=1 if constructed in north-east region of USA), `CT` (=1 if cooling tower used), `BW` (=1 if nuclear steam supply system manufactured by Babcock–Wilcox), `N` (cumulative number of power plants constructed by each architect-engineer), `PT` (=1 if partial turnkey plant).

	cost	date	T_1	T_2	capacity	PR	NE	CT	BW	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
⋮											
32	270.71	67.83	7	80	886	1	0	0	1	11	1

	Full model		Backward		Forward	
	Est	t	Est	t	Est	t
Int.	-14.24	-3.37	-13.26	-4.22	-7.62	-2.66
`date`	0.2	3.21	0.21	4.91	0.13	3.38
`logT1`	0.092	0.38				
`logT2`	0.29	1.05				
`logcap`	0.694	5.10	0.72	6.09	0.67	4.75
`PR`	-0.092	-1.20				
`NE`	0.25	3.35	0.24	3.36		
`CT`	0.12	1.82	0.14			
`BW`	0.033	0.33				
`log(N)`	-0.08	-1.74	-0.08	-2.11		
`PT`	-0.22	-1.83	-0.22	-1.99	-0.49	-4.77
s (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Recall: $\hat{y}$ is projection of y onto $\mathcal{M}(X)$

$\hookrightarrow$ Adding more variables (columns) into X "enlarges" $\mathcal{M}(X)$

. . . IF the rank increases by the # of new variables

Consider two extremes

- Adding a new variable $X_{p+1} \in \mathcal{M}^{\perp}(X)$
  - $\hookrightarrow$ Gives us completely "new" information.
- Adding a new variable $X_{p+1} \in \mathcal{M}(X)$
  - $\hookrightarrow$ Gives no "new" information — cannot even do least squares (why not?)

What if we are between the two extremes? What if

$$X_{p+1} \notin \mathcal{M}(X) \quad \text{but} \quad X(X^\top X)^{-1}X^\top X_{p+1} = HX_{p+1} \simeq X_{p+1}?$$

We can certainly fit the regression, but what will happen?

Using block matrix properties, have

$$\text{var}(\hat{\beta}) = \sigma^2 \left[(X \;\; X_{p+1})^\top (X \;\; X_{p+1}) \right]^{-1}$$

with

$$\left[(X \;\; X_{p+1})^\top (X \;\; X_{p+1}) \right]^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where

$$\begin{aligned}
A &= (X^\top X)^{-1} + (X^\top X)^{-1} X^\top X_{p+1} \\
&\qquad \times (X_{p+1}^\top X_{p+1} - X_{p+1}^\top H X_{p+1})^{-1} X_{p+1}^\top X (X^\top X)^{-1}, \\
B &= -(X^\top X)^{-1} X^\top X_{p+1} (X_{p+1}^\top X_{p+1} - X_{p+1}^\top H X_{p+1})^{-1}, \\
C &= -(X_{p+1}^\top X_{p+1} - X_{p+1}^\top H X_{p+1})^{-1} X_{p+1}^\top X (X^\top X)^{-1}, \\
D &= (X_{p+1}^\top X_{p+1} - X_{p+1}^\top H X_{p+1})^{-1}.
\end{aligned}$$

Multicollinearity: when p explanatories concentrate around a subspace of dimension $q < p$

[simplest case: pairs of variables that are correlated]

BUT: might exist even if pairs of variables appear uncorrelated!

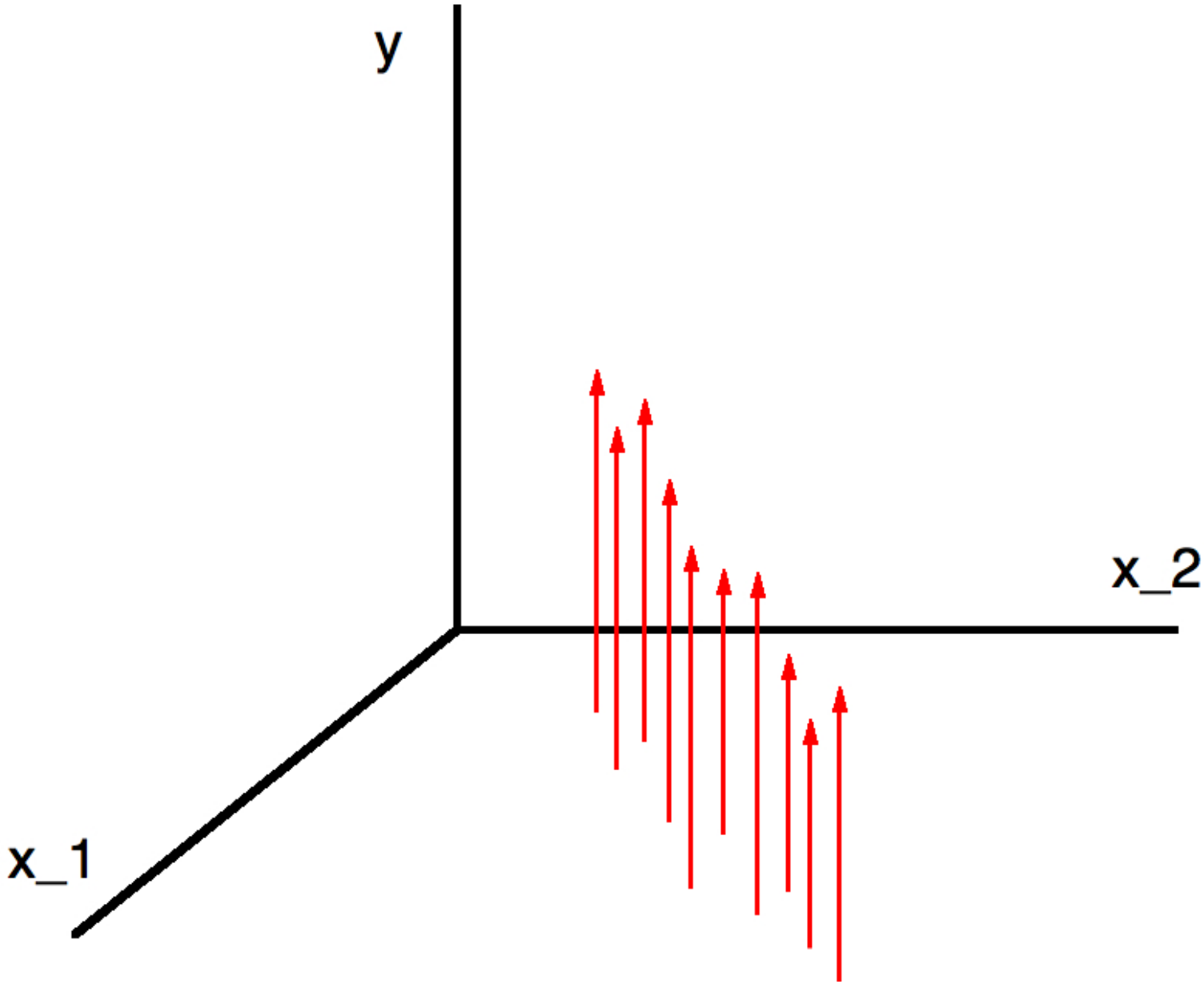
Can be caused by:

- Poor design [can try designing again],
- Inherent relationships [other remedies needed].

So what are the results?

- Huge variances of the estimators!
  - ↪ Can even flip signs for different data, to give the impression of inverse effects.
- Individual coefficients insignificant:
  - ↪ t-test p-values inflated.
- But global F-test might give significant result!

The Picket-Fence (Hocking & Pendleton)



Simple first steps:

- Look at scatterplots,
- Look at correlation matrix of explanatories,

Might not reveal more complex linear constraints, though.

- Look at the *variance inflation factors*:

$$VIF_j = \frac{\text{var}(\hat{\beta}_j)\|X_j\|^2}{\sigma^2} = \|X_j\|^2 \left[(X^\top X)^{-1}\right]_{jj} = \left[X^\top X\right]_{jj} \left[(X^\top X)^{-1}\right]_{jj}.$$

Can show that

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination for the regression

$$X_j = \beta_{0,j} + \beta_{1,j} X_1 + \cdots + \beta_{j-1,j} X_{j-1} + \beta_{j+1,j} X_{j+1} + \cdots + \beta_{p,j} X_p + \varepsilon,$$

measuring linear dependence of X_j on the other columns of X.

Let X_{-j} be the design matrix without the j-th variable. Then

$$R_j^2 = \frac{\|X_{-j}(X_{-j}^\top X_{-j})^{-1} X_{-j}^\top X_j\|^2}{\|X_j\|^2} \in [0,1]$$

is close to 1 if $\underbrace{X_{-j}(X_{-j}^\top X_{-j})^{-1} X_{-j}}_{H_{-j}} X_j \simeq X_j$.

Large values of VIF_j indicate that X_j is linearly dependent on the other columns of the design matrix.

Interpretation: how much the variance is inflated when including variable j as compared to the variance we would obtain if X_j were orthogonal to the other variables—how much worse are we doing as compared to the ideal case.

Rule of thumb: $VIF_j > 5$ or $VIF_j > 10$ considered to be "large".

Consider the spectral decomposition of $X^\top X$, $X^\top X = U \Lambda U^\top$ with $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$ and $U^\top U = I$. Then

$$\text{rank}(X^\top X) = \#\{j : \lambda_j \neq 0\}, \qquad \det(X^\top X) = \prod_{j=1}^{p} \lambda_j .$$

Hence "small" λ_j's mean "almost" reduced rank, revealing the effect of collinearity. Measure using *condition index*:

$$CI_j(X^\top X) := \sqrt{\lambda_{\text{max}}/\lambda_j}$$

Global "instability" measured by the *condition number*,

$$CN(X^\top X) = \sqrt{\lambda_{\text{max}}/\lambda_{\text{min}}}$$

Rule of thumb: $CN > 30$ indicates moderate to significant collinearity, $CN > 100$ indicates severe collinearity (choices vary).

If design faulty, may redesign.

$\hookrightarrow$ Otherwise? Inherent relationships between explanatories.

- Variable deletion - attempt to remove problematic variables
  - $\rightarrow$ E.g., by backward elimination.

- Choose an orthogonal basis for $\mathcal{M}(X)$ and use its elements as explanatories
  - $\rightarrow$ Use columns of U from spectrum, $X^\top X = U \Lambda U^\top$
  - $\rightarrow$ OK for prediction
  - $\rightarrow$ Problem: lose interpretability

Other approaches?

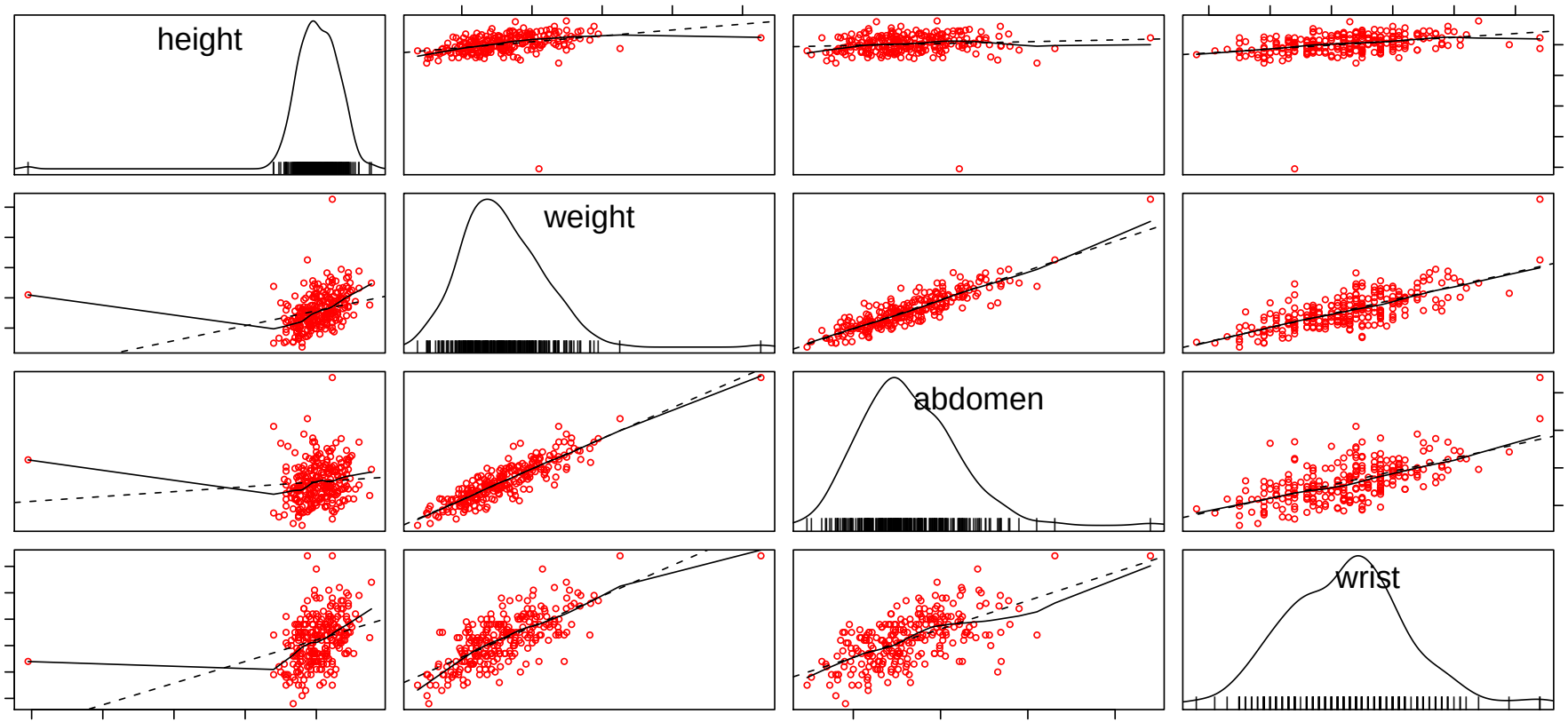
Body fat is measure of health $\rightarrow$ not easy to measure!
Collect 252 measurements on body fat and some explanatory variables.

Can we use measuring tape and scales only to find body fat?

Explanatory variables:

- age
- weight
- height
- biceps
- neck
- chest
- abdomen
- forearm
- hip
- thigh
- knee a
- ankle
- wrist

Some Scatterplots [library(car);scatterplot.matrix(. . .)]



Looks like we're in trouble. Let's go ahead and fit anyway . . .

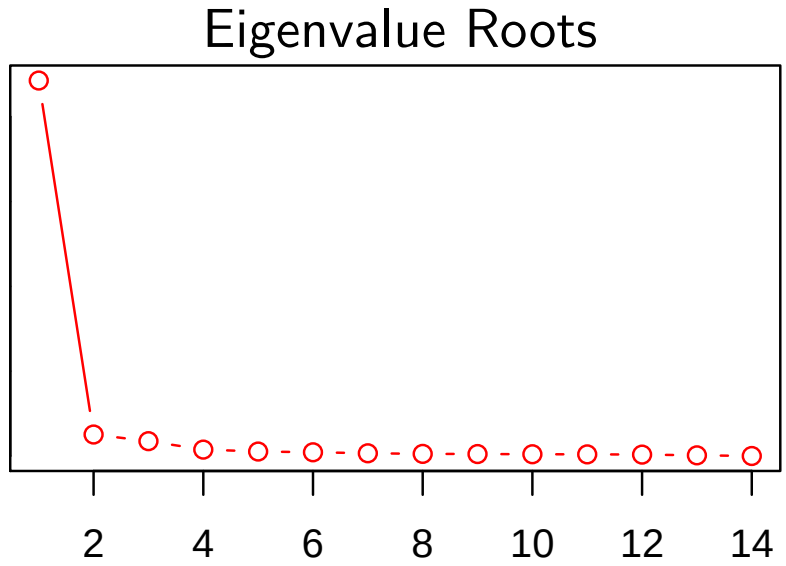
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−18.1885	17.3486	−1.05	0.2955
age	0.0621	0.0323	1.92	0.0562
weight	−0.0884	0.0535	−1.65	0.0998
height	−0.0696	0.0960	−0.72	0.4693
neck	−0.4706	0.2325	−2.02	0.0440
chest	−0.0239	0.0991	−0.24	0.8100
abdomen	0.9548	0.0864	11.04	0.0000
hip	−0.2075	0.1459	−1.42	0.1562
thigh	0.2361	0.1444	1.64	0.1033
knee	0.0153	0.2420	0.06	0.9497
ankle	0.1740	0.2215	0.79	0.4329
biceps	0.1816	0.1711	1.06	0.2897
forearm	0.4520	0.1991	2.27	0.0241
wrist	−1.6206	0.5349	−3.03	0.0027

$R^2 = 0.749$, F-test: $p < 2.2 \times 10^{-16}$.

	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
(Intercept)	−32.6564	0.1393	−1.2221	0.9730
age	0.1048	0.0153	0.0256	0.6252
weight	−0.1285	0.0502	−0.0237	0.8223
height	−0.0666	0.5207	−0.1005	0.7284
neck	−0.5086	0.0721	−0.4619	0.2635
chest	0.0168	0.9002	−0.0910	0.5877
abdomen	0.9750	0.0000	0.8924	0.0000
hip	−0.2891	0.1265	−0.0265	0.9130
thigh	0.3850	0.0565	0.0334	0.8793
knee	0.2218	0.5111	−0.1310	0.7366
ankle	0.4377	0.0694	−0.5037	0.3516
biceps	−0.1297	0.5485	0.4458	0.1179
forearm	0.8871	0.0174	0.2247	0.3750
wrist	−1.7378	0.0309	−1.5902	0.0560

	VIF
age	2.25
weight	33.51
height	1.67
neck	4.32
chest	9.46
abdomen	11.77
hip	14.80
thigh	7.78
knee	4.61
ankle	1.91
biceps	3.62
forearm	2.19
wrist	3.38

	CI
1	1.00
2	17.47
3	25.30
4	58.61
5	83.59
6	100.63
7	137.90
8	175.29
9	192.62
10	213.01
11	228.16
12	268.21
13	555.67



Condition Number $\simeq$ 556 !

```
 Multiple R-Squared:  0.7466,
F-statistic p-value:  < 2.2e-16
```

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(Intercept)	−22.6564	11.7139	−1.93	0.0543	
age	0.0658	0.0308	2.14	0.0336	2.05
weight	−0.0899	0.0399	−2.25	0.0252	18.82
neck	−0.4666	0.2246	−2.08	0.0388	4.08
abdomen	0.9448	0.0719	13.13	0.0000	8.23
hip	−0.1954	0.1385	−1.41	0.1594	13.47
thigh	0.3024	0.1290	2.34	0.0199	6.28
forearm	0.5157	0.1863	2.77	0.0061	1.94
wrist	−1.5367	0.5094	−3.02	0.0028	3.09

Define $Z = XU$ as design matrix. R^2=0.749, F-test p-value<2.2×10^{-16}

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−18.1885	17.3486	−1.05	0.2955
Z[, 1]	−0.1353	0.0619	−2.19	0.0297
Z[, 2]	−0.0168	0.0916	−0.18	0.8546
Z[, 3]	0.2372	0.1070	2.22	0.0276
Z[, 4]	−0.7188	0.0571	−12.58	0.0000
Z[, 5]	0.0248	0.0827	0.30	0.7649
Z[, 6]	0.4546	0.1001	4.54	0.0000
Z[, 7]	0.5903	0.1366	4.32	0.0000
Z[, 8]	−0.1207	0.1742	−0.69	0.4890
Z[, 9]	−0.0836	0.1914	−0.44	0.6627
Z[, 10]	0.5043	0.2082	2.42	0.0162
Z[, 11]	−0.5735	0.2254	−2.54	0.0116
Z[, 12]	0.3007	0.2628	1.14	0.2536
Z[, 13]	1.5168	0.5447	2.78	0.0058

- Eigenvector approach rotates space so as to "free" the dependence of one coefficient β_j on others $\{\beta_i\}_{i \neq j}$

  ↪ Imposes constraint on X (orthogonal columns)

Problem: lose interpretability! (prediction OK)

  - Example: most significant "rotated" term in fat data: `Z[,4]=-0.01*age -0.058*weight -0.011*height +0.46*neck -0.144*chest -0.441*abdomen +0.586*hip +0.22*thigh -0.197*knee -0.044*ankle -0.07*biceps -0.33*forearm -0.249*wrist`

- Other approach to reduce this strong dependence?

  ↪ Impose constraint on β! How? (introduces bias)

Multicollinearity problem is that $\det\left[(X^\top X)^{-1}\right] \approx 0$
[i.e. $X^\top X$ *almost not invertible*]

A Solution: add a "small amount" of a full rank matrix to $X^\top X$.

For reasons to become clear soon, we *standardise* the design matrix:

- Write $X = (\mathbf{1}\ W)$, $\beta = (\beta_0\ \gamma)^\top$
- Recentre/rescale the covariates defining: $Z_j = \frac{\sqrt{n}}{\mathrm{sd}(W_j)}(W_j - \mathbf{1}\overline{W}_j)$
  - ↪ Coefficients now have common scale
  - ↪ Interpretation of β_j slightly different: not "mean impact on response per unit change of explanatory variable", but now "mean impact on response per unit deviation of explanatory variable from its mean, measured in units of standard deviation"
- The Z_j are all orthogonal to $\mathbf{1}$ and are of unit norm.

- Since $Z_j \perp \mathbf{1}$ for all, j, we can estimate β_0 and γ by two separate regressions (orthogonality).
- Least squares estimators become

$$\hat{\beta}_0 = \overline{Y}, \quad \hat{\gamma} = (Z^\top Z)^{-1} Z^\top Y.$$

- Ridge regression replaces $Z^\top Z$ by $Z^\top Z + \lambda I$ (i.e. adds a "ridge")

$$\boxed{\hat{\beta}_0 = \overline{Y}, \quad \hat{\gamma} = (Z^\top Z + \lambda I)^{-1} Z^\top Y}.$$

Adding λI to $Z^\top Z$ makes inversion more stable
$\hookrightarrow$ λ called *ridge parameter*.

$\rightarrow$ Ridge term λI seems slightly ad-hoc. Motivation?

$\hookrightarrow$ Can see that $(\hat{\beta}_0 \quad \hat{\gamma}) = (\overline{Y} \quad (Z^\top Z + \lambda I)^{-1} Z^\top Y)$ minimizes

$$\| Y - \beta_0 \mathbf{1} - Z\gamma \|_2^2 + \lambda \|\gamma\|_2^2$$

or equivalently

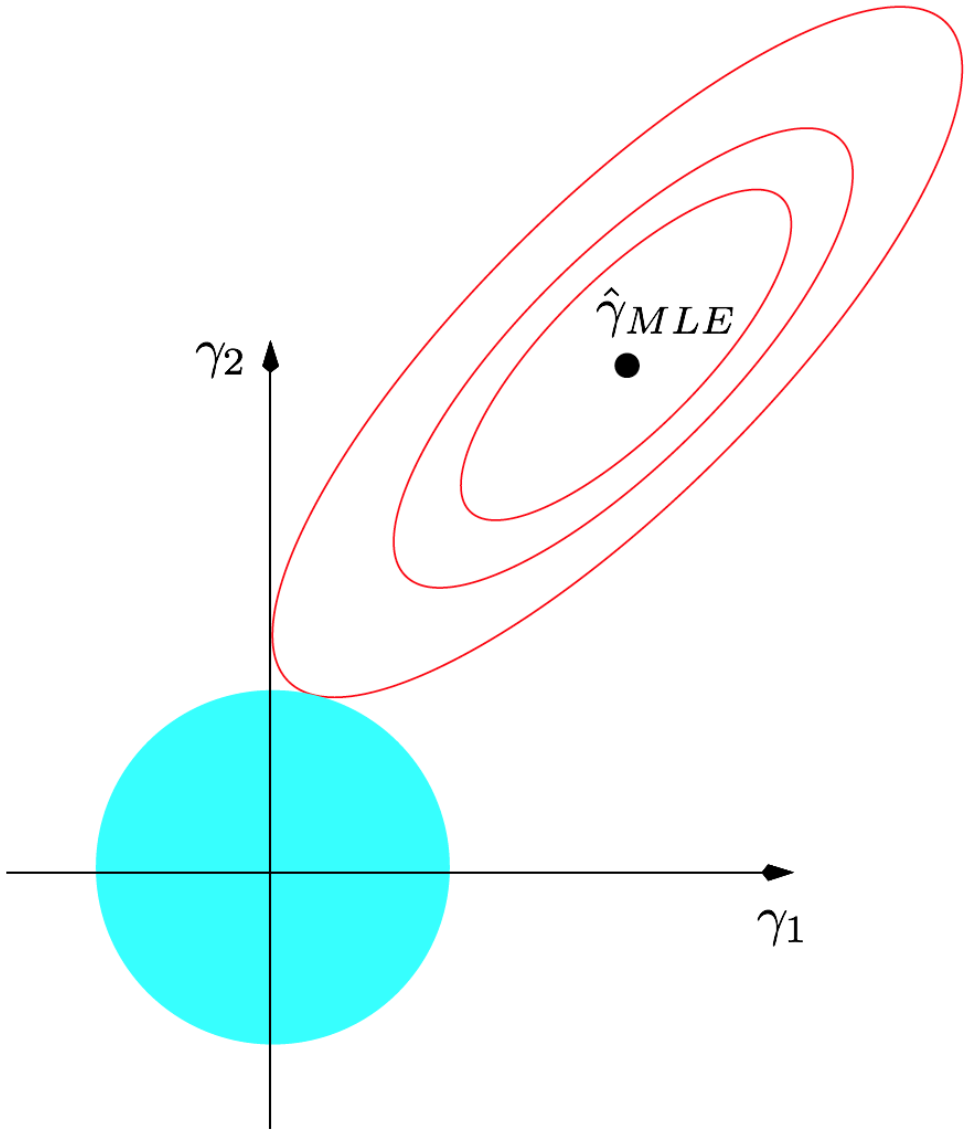
$$\| Y - \beta_0 \mathbf{1} - Z\gamma \|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} \gamma_j^2 = \|\gamma\|_2^2 \leq r(\lambda)$$

instead of least squares estimator which minimizes

$$\| Y - \beta_0 \mathbf{1} - Z\gamma \|_2^2.$$

Idea: in the presence of collinearity, coefficients are ill-defined: a wildly positive coefficient can be cancelled out by a largely negative coefficient (many coefficient combinations can produce the same effect). By imposing a *size* constraint, we limit the possible coefficient combinations!

L^2 Shrinkage [Ridge Regression]



Proposition

Let $Z_{n\times q}$ be a matrix of rank $r \leq q$ with centred column vectors of unit norm. Given $\lambda > 0$, the unique minimiser of

$$Q(\hat{\beta}_0, \hat{\gamma}) = \|y - \hat{\beta}_0 \mathbf{1} - Z\hat{\gamma}\|_2^2 + \lambda\|\hat{\gamma}\|_2^2$$

is

$$(\hat{\beta}_0\,,\,\hat{\gamma}) = (\overline{y}\,,\,(Z^\top Z + \lambda I)^{-1} Z^\top y).$$

Proof.

Write

$$y = \underbrace{(y - \bar{y}\mathbf{1})}_{=y^* \in \mathcal{M}^\perp(\mathbf{1})} + \underbrace{\bar{y}\mathbf{1}}_{\in \mathcal{M}(\mathbf{1})}$$

Note also that by assumption $\mathbf{1} \in \mathcal{M}^\perp(Z)$. Therefore by Pythagoras' theorem

$$\|y - \hat{\beta}_0\mathbf{1} - Z\hat{\gamma}\|_2^2 = \|\underbrace{(\bar{y} - \hat{\beta}_0)\mathbf{1}}_{\in \mathcal{M}(\mathbf{1})} + \underbrace{(y^* - Z\hat{\gamma})}_{\in \mathcal{M}^\perp(\mathbf{1})}\|_2^2 = \|(\bar{y} - \hat{\beta}_0)\mathbf{1}\|_2^2 + \|(y^* - Z\hat{\gamma})\|_2^2.$$

Therefore, $\min_{\hat{\beta}_0, \hat{\gamma}} Q(\hat{\beta}_0, \hat{\gamma}) = \min_{\hat{\beta}_0} ||(\bar{y} - \hat{\beta}_0)\mathbf{1}||_2^2 + \min_{\hat{\gamma}} \left\{ ||(y^* - Z\hat{\gamma})||_2^2 + \lambda ||\hat{\gamma}||_2^2 \right\}$

Clearly, $\arg\min_{\hat{\beta}_0} ||(\bar{y} - \hat{\beta}_0)\mathbf{1}||_2^2 = \bar{y}$ while the second component can be written

$$\min_{\hat{\gamma} \in \mathbb{R}^q} \left\| \begin{pmatrix} Z \\ {\scriptstyle n \times q} \\ \sqrt{\lambda} I_{q \times q} \end{pmatrix} \hat{\gamma} - \begin{pmatrix} y^* \\ {\scriptstyle n \times 1} \\ \mathbf{0}_{q \times 1} \end{pmatrix} \right\|_2^2$$

using block notation. This is the usual least squares problem with solution

$$\left[(Z^\top, \sqrt{\lambda} I_{q \times q}) \begin{pmatrix} Z \\ \sqrt{\lambda} I_{q \times q} \end{pmatrix} \right]^{-1} (Z^\top, \sqrt{\lambda} I_{q \times q}) \begin{pmatrix} y^* \\ \mathbf{0}_{q \times 1} \end{pmatrix} = (Z^\top Z + \lambda I)^{-1} Z^\top y^*$$

Note that $Z^\top Z + \lambda I$ is indeed invertible. Writing $Z^\top Z = U \Lambda U^\top$, we have

$$Z^\top Z + \lambda I = U \Lambda U^\top + U(\lambda I_{q \times q}) U^\top = U(\Lambda + \lambda I_{q \times q}) U^\top$$

and $\Lambda = \text{diag}\{\underbrace{\lambda_1, \ldots, \lambda_r}_{>0}, \underbrace{\lambda_{r+1}, \ldots, \lambda_q}_{=0}\}$ ($Z^\top Z \succeq 0$ & $rank(Z^\top Z) = rank(Z)$).

To complete the proof, observe that $Z^\top y^* = Z^\top y - \bar{y} Z^\top \mathbf{1} = Z^\top y$. □

Note that if the SVD of Z is $Z = V\Omega U^\top$, last steps of previous proof may be used to show that

$$\hat{\gamma} = \sum_{j=1}^{q} \frac{\omega_j}{\omega_j^2 + \lambda}(v_j^\top y)u_j,$$

where the v_js and u_js are the columns of V and U, respectively.

Compare this to the ordinary least squares solution, when $\lambda = 0$:

$$\hat{\gamma} = \sum_{j=1}^{q} \frac{1}{\omega_j}(v_j^\top y)u_j,$$

which is not even defined if Z is of reduced rank.

Role of λ is to reduce the size of $1/\omega_j$ when ω_j becomes very small.

Proposition

Let $\hat{\gamma}$ be the ridge regression estimator of γ. Then

$$bias(\hat{\gamma}, \gamma) = -\lambda \left(Z^\top Z + \lambda I_q\right)^{-1} \gamma$$

and

$$cov(\hat{\gamma}) = \sigma^2 (Z^\top Z + \lambda I)^{-1} Z^\top Z (Z^\top Z + \lambda I)^{-1}.$$

Proof.

Since $\mathbb{E}(\hat{\gamma}) = (Z^\top Z + \lambda I)^{-1} Z^\top \mathbb{E}(y) = (Z^\top Z + \lambda I)^{-1} Z^\top Z\gamma$, the bias is

$$\begin{aligned}
\text{bias}(\hat{\gamma}, \gamma) &= \mathbb{E}(\hat{\gamma}) - \gamma = \{(Z^\top Z + \lambda I)^{-1} Z^\top Z - I\}\gamma \\
&= \left\{ \left(\frac{1}{\lambda} Z^\top Z + I\right)^{-1} \left(\frac{1}{\lambda} Z^\top Z + I - I\right) - I \right\} \gamma \\
&= \left\{ I - \left(\frac{1}{\lambda} Z^\top Z + I\right)^{-1} - I \right\} \gamma = -\left(\frac{1}{\lambda} Z^\top Z + I\right)^{-1} \gamma.
\end{aligned}$$

The covariance term is obvious. □

Corollary

Assume that $\mathrm{rank}(Z_{n\times q}) = q$ *and let*

$$\tilde{\gamma} = (Z^\top Z)^{-1} Z^\top y \qquad \& \qquad \hat{\gamma}_\lambda = (Z^\top Z + \lambda I)^{-1} Z^\top y$$

be the least squares estimator and ridge estimator, respectively. Then,

$$\mathbb{E}\left\{(\tilde{\gamma} - \gamma)(\tilde{\gamma} - \gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda - \gamma)(\hat{\gamma}_\lambda - \gamma)^\top\right\} \succeq 0$$

for all $\lambda \in (0, 2\sigma^2/\|\gamma\|^2)$.

Ridge estimator uniformly better than least squares! How can this be?
(What happened to Gauss-Markov?)

- Gauss-Markov only covers unbiased estimators – but ridge estimator biased.
- A bit of bias can improve the MSE by reducing variance.
- Also, there is a catch: the "right" range for λ depends on unknowns.
- Choosing a good λ is all about balancing bias and variance.

Proof.

From our bias/variance calculations on the ridge estimator, we have

$$\mathbb{E}\left\{(\tilde{\gamma}-\gamma)(\tilde{\gamma}-\gamma)^\top\right\} - \mathbb{E}\left\{(\hat{\gamma}_\lambda-\gamma)(\hat{\gamma}_\lambda-\gamma)^\top\right\} =$$

$$\sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - (\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\sigma^2\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1} - \lambda^2\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I}\right)^{-1}\gamma\gamma^\top\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I}\right)^{-1}$$

$$= \lambda(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\left(\sigma^2(2\boldsymbol{I}+\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}) - \lambda\gamma\gamma^\top\right)(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}.$$

To go from 2nd to 3rd line, we wrote

$$\sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} = \sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}$$

$$= (\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}(\sigma^2\boldsymbol{Z}^\top\boldsymbol{Z}+2\sigma^2\lambda\boldsymbol{I}+\sigma^2\lambda^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1})(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}$$

and did the tedious (but straighforward) algebra. The final term can be made positive definite if

$$2\sigma^2\boldsymbol{I} + \sigma^2\lambda(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} - \lambda\gamma\gamma^\top \succeq 0.$$

Noting that we can always normalise/complete γ to an ONB $\{\gamma/||\gamma||, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{q-1}\}$ of $\mathbb{R}^q$ we may write

$$\boldsymbol{I} = \frac{\gamma\gamma^\top}{||\gamma||^2} + \sum_{j=1}^{q-1}\boldsymbol{\theta}_j\boldsymbol{\theta}_j^\top.$$

We thus see that $\lambda \in (0, 2\sigma^2/||\gamma||^2)$ suffices for positive definiteness to hold true.

□

Role of λ: Regulates Bias–Variance tradeoff

- $\lambda \uparrow$ decreases variance (e.g. due to collinearity) but increases bias
- $\lambda \downarrow$ decreases bias but variance inflated if collinearity exists

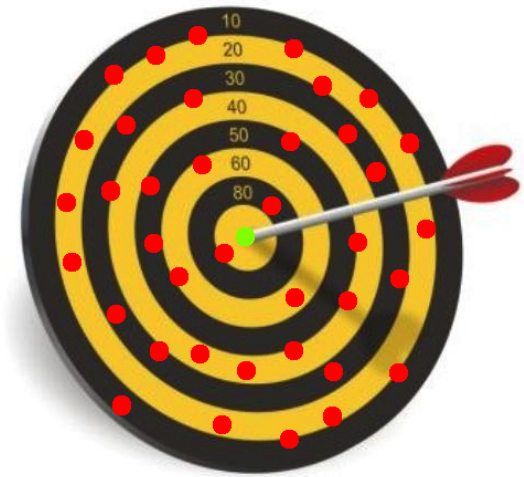
Recall:

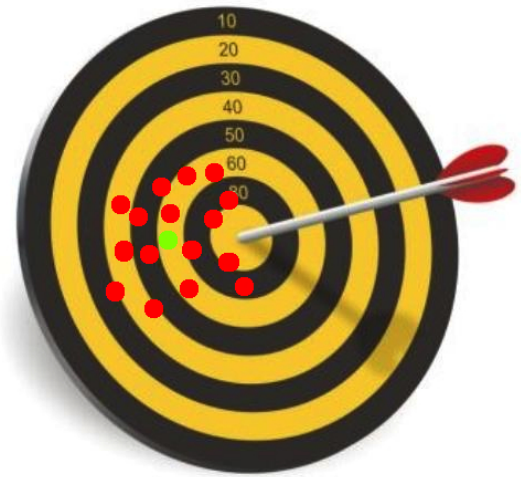
$$\mathbb{E}\|\hat{\gamma} - \gamma\|^2 = \underbrace{\mathbb{E}\|\hat{\gamma} - \mathbb{E}\hat{\gamma}\|^2}_{Variance = trace[cov(\hat{\gamma})]} + \underbrace{\|\mathbb{E}\hat{\gamma} - \gamma\|^2}_{Bias^2} + \underbrace{2(\mathbb{E}\hat{\gamma} - \gamma)^\top \mathbb{E}[\hat{\gamma} - \mathbb{E}\hat{\gamma}]}_{=0}$$

Note that if $Z^\top Z = U \Omega U^\top$, then $trace(cov(\hat{\gamma})) = \sum_{j=1}^{q} \frac{\omega_i}{\omega_i^2 + \lambda} \sigma^2$

So choose λ so as to optimally increase bias/decrease variance

Use cross validation!





Motivated from Ridge Regression formulation can consider:

$$\text{min!} \qquad \|Y - \beta_0 \mathbf{1} - Z\gamma\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} |\gamma_j| = \|\gamma\|_1 \leq r(\lambda)$$

$$\Longleftrightarrow$$

$$\text{min!} \qquad \|Y - \beta_0 \mathbf{1} - Z\gamma\|_2^2 + \lambda\|\gamma\|_1.$$

Shrinks coefficient *size* by different version of *magnitude*.

- Resulting estimator non–linear in Y
- No explicit form available (unless $Z^\top Z = I$), needs quadratic programming algorithm

Why choose a different type of norm?
Because L^1 penalty (almost) produces a "continuous" model selection!

When the explanatory variables are orthogonal (i.e. $Z^\top Z = I$), then the LASSO exactly performs model selection via thresholding:

Theorem

Consider the linear model

$$\underset{n\times 1}{Y} = \underset{1\times 1}{\beta_0}\,\underset{n\times 1}{\mathbf{1}} + \underset{n\times(p-1)}{Z}\,\underset{(p-1)\times 1}{\gamma} + \underset{n\times 1}{\varepsilon}$$

where $Z^\top \mathbf{1} = 0$ *and* $Z^\top Z = I$. *Let* $\hat{\gamma}$ *be the least squares estimator of* γ,

$$\hat{\gamma} = (Z^\top Z)^{-1} Z^\top Y = Z^\top Y.$$

Then, the unique solution to the LASSO problem

$$\min_{\beta_0\in\mathbb{R},\gamma\in\mathbb{R}^{p-1}} \left\{ || Y - \beta_0 \mathbf{1} - Z\gamma||_2^2 + \lambda||\gamma||_1 \right\}$$

is given by $(\hat{\beta}_0, \check{\gamma}) = (\beta_0, \check{\gamma}_1, \ldots, \check{\gamma}_{p-1})$, *defined as*

$$\hat{\beta}_0 = \bar{Y} \qquad \& \qquad \check{\gamma}_i = sgn(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \frac{\lambda}{2}\right)_+, \quad i = 1, ..., p-1.$$

Proof.

Note that since $Z^\top \mathbf{1} = 0$ and since β_0 does not appear in the L^1 penalty, we have

$$\hat{\beta}_0 = (\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1} Y = \bar{Y}.$$

Thus, the LASSO problem reduces to

$$\min_{\beta_0 \in \mathbb{R}, \gamma \in \mathbb{R}^{p-1}} \left\{ \| Y - \beta_0 \mathbf{1} - Z\gamma \|_2^2 + \lambda \|\gamma\|_1 \right\} = \min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \|u - Z\gamma\|_2^2 + \lambda \|\gamma\|_1 \right\}.$$

where $u = Y - \bar{Y}\mathbf{1}$ for tidiness. Expanding the squared norm gives

$$\|u - Z\hat{\gamma}\|_2^2 = u^\top u - 2u^\top Z\gamma + \gamma^\top \underbrace{(Z^\top Z)}_{=I} \gamma = u^\top u - 2\underbrace{Y^\top Z}_{=\hat{\gamma}^\top}\gamma + 2\bar{Y}\underbrace{\mathbf{1}^\top Z}_{=0}\gamma + \gamma^\top \gamma$$

Since $u^\top u$ does not depend on γ, we see that the LASSO objective function is

$$-2\hat{\gamma}^\top \gamma + \|\gamma\|_2^2 + \lambda \|\gamma\|_1.$$

Clearly, this has the same minimizer if multiplied across by $1/2$, i.e.

$$-\hat{\gamma}^\top \gamma + \tfrac{1}{2}\|\gamma\|_2^2 + \tfrac{1}{2}\lambda \|\gamma\|_1 = \textstyle\sum_{i=1}^{p-1} \left(-\hat{\gamma}_i \gamma_i + \tfrac{1}{2}\gamma_i^2 + \tfrac{\lambda}{2}|\gamma_i| \right).$$

Notice that we now have a sum of $p-1$ objective functions, each depending only on one γ_i. We can thus optimise each separately. That is, for any given $i \leq p-1$, we must minimise

$$-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|.$$

We distinguish 3 cases:

1. Case $\hat{\gamma}_i = 0$. In this case, the objective function becomes $\frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ and it is clear that it is minimised when $\gamma_i = 0$. So in this case $\check{\gamma}_i = 0$.
2. Case $\hat{\gamma}_i > 0$. In this case, the objective function $-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in [0, \infty)$ because the term $-\hat{\gamma}_i \gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \geq 0$, the objective function becomes

$$-\hat{\gamma}_i \gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}\gamma_i = \left(\frac{\lambda}{2} - \hat{\gamma}_i\right)\gamma_i + \frac{1}{2}\gamma_i^2.$$

   If $\frac{\lambda}{2} - \hat{\gamma}_i \geq 0$, then the minimum over $\gamma_i \in [0, \infty)$ is clearly at $\gamma_i = 0$. Otherwise, when $\frac{\lambda}{2} - \hat{\gamma}_i < 0$, we differentiate and find the minimum at $\gamma_i = \hat{\gamma}_i - \lambda/2 > 0$. In summary, $\check{\gamma}_i = (\hat{\gamma}_i - \lambda/2)_+ = sgn(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

3. Case $\hat{\gamma}_i < 0$. In this case, the objective function $-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}|\gamma_i|$ is minimised somewhere in the range $\gamma_i \in (-\infty, 0]$ because the term $-\hat{\gamma}_i\gamma_i$ is negative there (and all other terms are positive). But when $\gamma_i \leq 0$, the objective function becomes

$$-\hat{\gamma}_i\gamma_i + \frac{1}{2}\gamma_i^2 + \frac{\lambda}{2}(-\gamma_i) = \left(\frac{\lambda}{2} + \hat{\gamma}_i\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2 = \left(\frac{\lambda}{2} - |\hat{\gamma}_i|\right)(-\gamma_i) + \frac{1}{2}\gamma_i^2.$$

If $\frac{\lambda}{2} - |\hat{\gamma}_i| \geq 0$, then the minimum over $\gamma_i \in (-\infty, 0]$ is clearly at $\gamma_i = 0$, since $-\gamma_i$ ranges over $[0, \infty)$. Otherwise, when $\frac{\lambda}{2} - |\hat{\gamma}_i| < 0$, we differentiate and find the minimum at $\gamma_i = -|\hat{\gamma}_i| + \lambda/2 < 0$, which we may re-write as:

$$-|\hat{\gamma}_i| + \lambda/2 = -\left(|\hat{\gamma}_i| - \lambda/2\right) = sgn(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \lambda/2\right).$$

In summary, $\check{\gamma}_i = sgn(\hat{\gamma}_i)(|\hat{\gamma}_i| - \lambda/2)_+$.

The proof is now complete, as we can see that all three cases yield

$$\check{\gamma}_i = sgn(\hat{\gamma}_i)\left(|\hat{\gamma}_i| - \frac{\lambda}{2}\right)_+, \quad i = 1, ..., p-1.$$

□

How can we interpret the LASSO in terms of ANOVA in the orthogonal case?

Corollary

In the context of the previous theorem, and assuming that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$*, model selection using the LASSO tuned by* $\lambda > 0$ *is equivalent to including only coefficients significant at level* $\alpha = 2(1 - G_{t_{n-p}}(\lambda/(2S)))$*, where* $G_{t_{n-p}}$ *is the CDF of Student's t-distribution.*

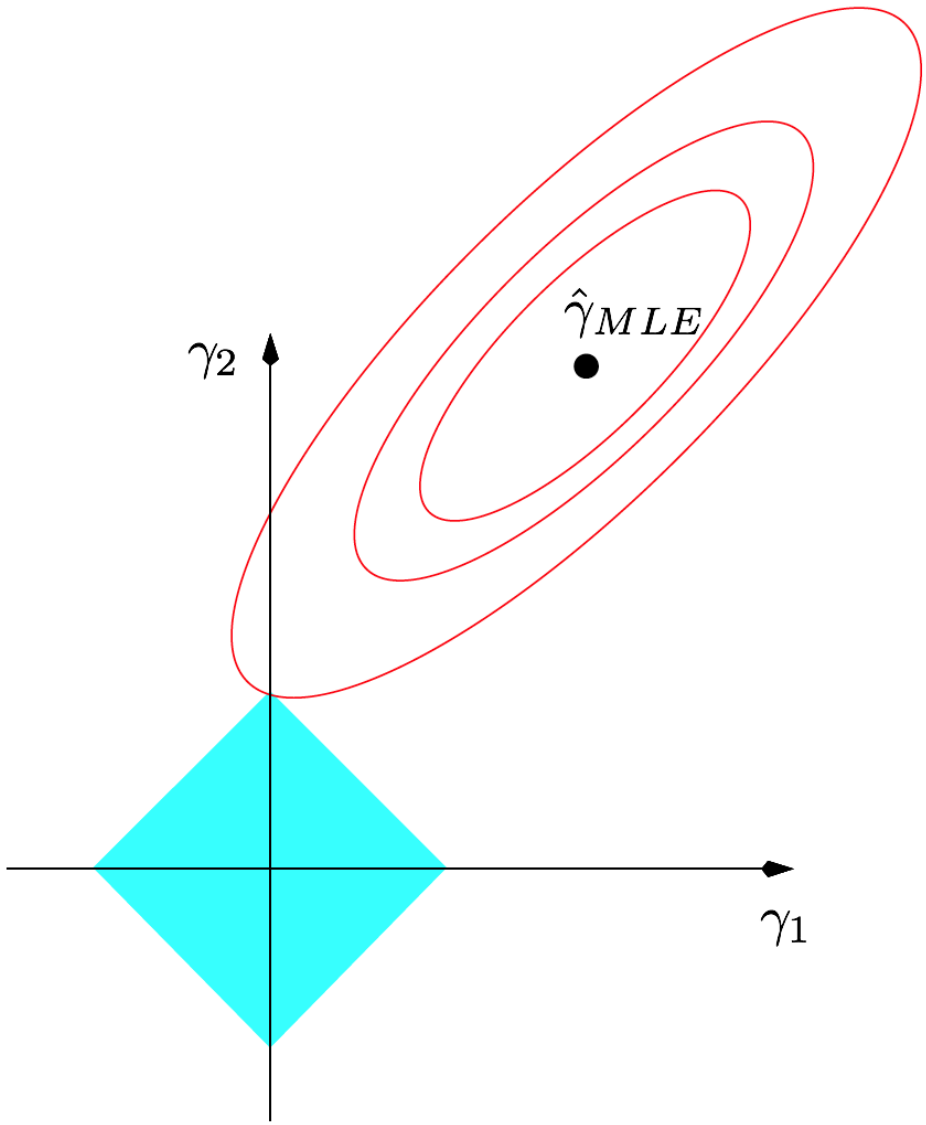
Proof.

Remember that a coefficient γ_j is pronounced statistically significant at level α if $\{H_0 : \gamma_j = 0\}$ is rejected at level α. Under the setting when $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, this happens when $|\hat{\gamma}_j| > t_{n-p}(1 - \alpha/2)S$. So equating

$$\frac{\lambda}{2} = t_{n-p}(1-\alpha/2)S \implies 1-\frac{\alpha}{2} = G_{t_{n-p}}(\lambda/(2S)) \implies \alpha = 2(1-G_{t_{n-p}}(\lambda/(2S)))$$

□

L^1 Shrinkage [The LASSO]



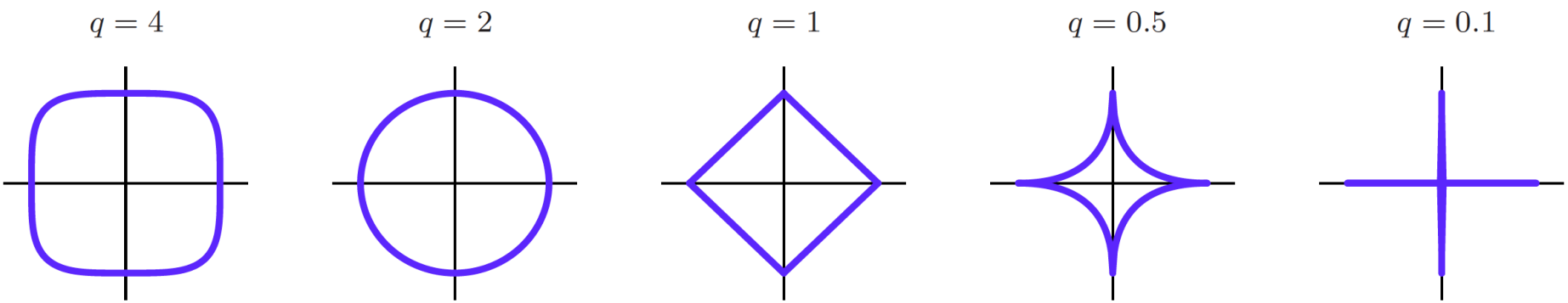
Intuition: L_1 norm induces "sharp" balls!

- Balls more concentrated around the axes
- Induces model selection by regulating the lasso (through λ)

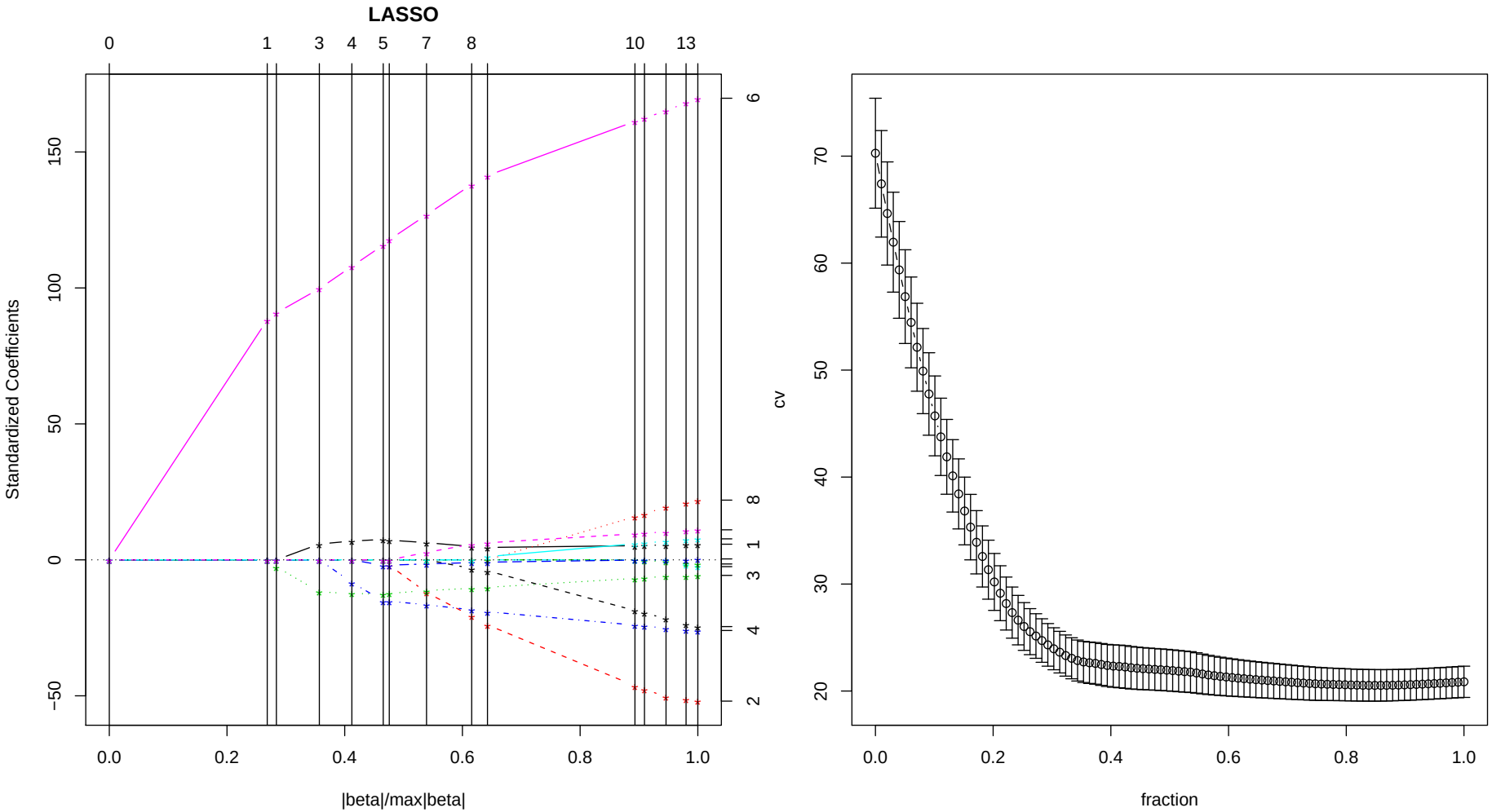
Extreme case: L^0 "Norm", gives best subsets selection!

$$\|\gamma\|_0 = \sum_{j=1}^{p-1} |\gamma_j|^0 = \sum_{j=1}^{p-1} \mathbf{1}_{\{\gamma_j \neq 0\}} = \#\{j : \gamma_j \neq 0\}$$

Generally: $\|\gamma\|_p^p = \sum_{j=1}^{p-1} |\gamma_j|^p$, sharp balls for $0 < p \leq 1$



LASSO and CV for different values of $r(\lambda)/\|\hat{\gamma}\|_1$

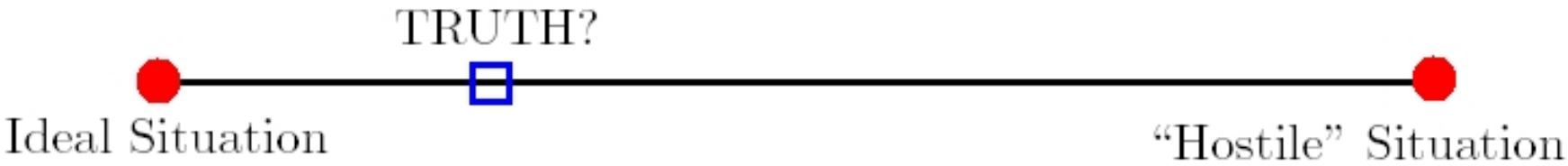


Robust Linear Modeling

The "success" of the LSE in a regression model depends on "assumptions":

- Normality (LSE optimal in this case)
- Not many "extreme" observations (LSE affected from "extremities")

Picture:



- **Resistant** procedure: not strongly affected by changes to data.
- **Robust** procedure: not strongly affected by departures from distribution.
  - Often: Robust $\Leftrightarrow$ Resistant

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} F$, estimate $\mu = \int_{-\infty}^{\infty} yF(dy)$ by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \gamma)^2$$

Some observations:

- Average $\bar{y}$ is optimal (MLE) when F is Normal.
- Extremely sensitive to outliers (low *breakdown point*).
- Can be made arbitrarily large by suitably perturbing a single sample value:

$$y_1 \mapsto y_1 + c \implies \bar{y} \mapsto \bar{y} + c/n.$$

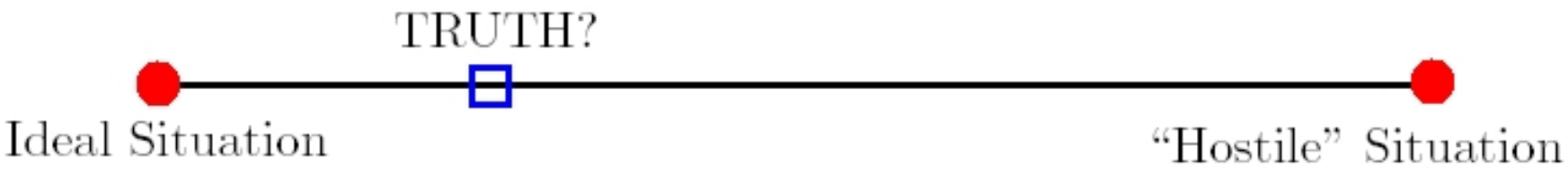
- If c large relative to $n \to$ adverse effects ...
- Also objective might not be optimal for other possible F's ...

Can we "cure" sensitivity by using different *distance function*? For instance, take

$$m = \arg\min_{\gamma \in \mathbb{R}} \sum_{i=1}^{n} |y_i - \gamma| = \begin{cases} y_{(k+1)}, & n = 2k+1, \\ \frac{y_{(k)} + y_{(k+1)}}{2}, & n = 2k. \end{cases}$$

- Median much less sensitive to bad values.
- Higher breakdown point: must blow up at least 50% of obs to blow m up.
- Median is optimal (MLE) when F is Laplace.
- But how well does m perform when $F \simeq$ Normal (*relative efficiency)*?

Remember picture:



Other alternatives? α-Trimmed mean: discards "most extreme" observations:

$$trm = \frac{1}{|E^c|} \sum_{i \notin E} y_i,$$

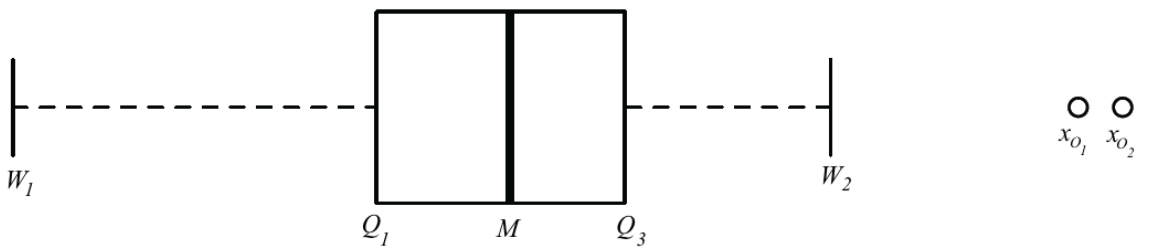
E being subset of $\alpha \times n$ most extreme observations from each end.

- When n is odd, using the median essentially takes E to be all but the most central observations.
- But we could also say, discard all the "outliers", based on Tukey's definition:
  - Take the first and third quartile, Q_1 and Q_3, and define the whiskers,

$$W_1 = \min_{1 \leq j \leq n} \{y_j : y_j \geq Q_1 - 1.5 \times |Q_3 - Q_1|\}$$

$$W_2 = \max_{1 \leq j \leq n} \{y_j : y_j \leq Q_3 + 1.5 \times |Q_3 - Q_1|\}$$

  - So that $E = \{i \in \{1, \ldots, n\} : y_i \notin [W_1, W_2]\}$ are the outliers.



Both m and trm may 'throw away' useful information. We can view them as special cases of the weighted mean:

$$wm = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}.$$

Weights downplaying "outliers", e.g. using Tukey's definition, you could choose

$$w_i = \begin{cases} 1 & \text{if } i \notin E, \\ \left(\frac{|y_i - m|}{|W_2 - W_1|} \right)^{-1} & \text{if } i \in E. \end{cases}$$

or some other "robust notion of dispersion".
Of course, you could also use a more flexible notion of what an "extreme" observation is, not based on boxplots, e.g. based on how large a "standardised" magnitude is, where the standardisation is done robustly. Would need to choose:

1. What we mean by magnitude/size
2. What we take as a location/centre (to re-centre)
3. What we take as a scale (to re-scale)

We want robust notions of location and scale to standardise because non-robust versions will be skewed/inflated on account of the outliers, making them seem as not so extreme post-standardisation. The problem appears cyclical – but in the 1D case, the fact that we can order observations (order statistics) simplifies things (as we saw in the Tukey case).

Note that the marginal case, is like a regression problem that only contains an intercept (no covariates).

More general regression situation is similar. Have:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim F, \quad \mathbb{E}[\varepsilon] = 0, \text{cov}[\varepsilon] = \sigma^2 I$$

LSE for β given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top y = \arg\min_{\gamma \in \mathbb{R}^p} \sum_{k=1}^{n} (y_i - x_i^\top \gamma)^2$$

- Optimal at $F =$ Normal
- Disastrous if $y_i \mapsto y_i + c$ with c large:

$$\hat{\beta} \mapsto \hat{\beta} + (X^T X)^{-1} x_i \, c$$

- Gauss-Markov: optimal *linear* for any F
  - ↪ May not be overall optimal for other F's

- L^1 regression: $\tilde{\beta} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_{k=1}^{n} |y_i - x_i^\top \gamma|$
- Trimmed least squares: $\check{\beta} = \arg\min_{\gamma \in \mathbb{R}^p} \sum_{i=1}^{K} (y_i - x_i^\top \gamma)^2_{(i)}$, where we set $K = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$
- Weighted least squares: $\check{\beta} = (X^\top V^{-1} X)^{-1} X^\top V^{-1} Y$ for a diagonal weight matrix V (recall earlier lecture):

$$V = \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{pmatrix}.$$

- Of course now we can't just a priori order observations to define the weights that would "downweight extremes" (comare to the 1D case).

Would like to **formalise** the concept of robust/resistant estimation

$\rightarrow$ Find a general formulation of which above are special cases.

$\rightarrow$ Find a fitting procedure that can overcome the lack of a priori ordering.

Seek a unifying approach:

- Instead of $(\cdot)^2$ or $|\cdot|$, consider a more general distance function $\rho(\cdot)$.

MLE when errors are Gaussian is obtained as maximising loglikelihood kernel

$$\hat{\beta} = \underset{\gamma \in \mathbb{R}^p}{\arg\max} - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - x_i^\top \gamma}{\sigma} \right)^2$$

Replacing $\rho(u) = u^2$ by general $\rho(\cdot)$ yields:

$$\widehat{\beta} := \underset{\gamma \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho \left(\frac{y_i - x_i^\top \gamma}{\sigma} \right)$$

Call this an **M**(aximum likelihood like)-**Estimator**.

Obtaining $\arg\min_{\gamma \in \mathbb{R}^p} \sum_{i=1}^n \rho\left(\frac{y_i - x_i^\top \gamma}{\sigma}\right)$ reduces to solving

$$\sum_{i=1}^n x_i^\top \psi\left(\frac{y_i - x_i^\top \gamma}{\sigma}\right) = 0$$

with $\psi(t) = d\rho(t)/dt$. Letting $w(u) = \psi(u)/u$ this reduces to

$$\sum_{i=1}^n w_i x_i^\top (y_i - x_i^\top \gamma) = 0, \quad \text{where } w_i = w\left(\frac{y_i - x_i^\top \gamma}{\sigma}\right).$$

But this is simply the weighting scenario!

- Robust Regression can be written as a Weighted Regression, but the weights depend on the data.

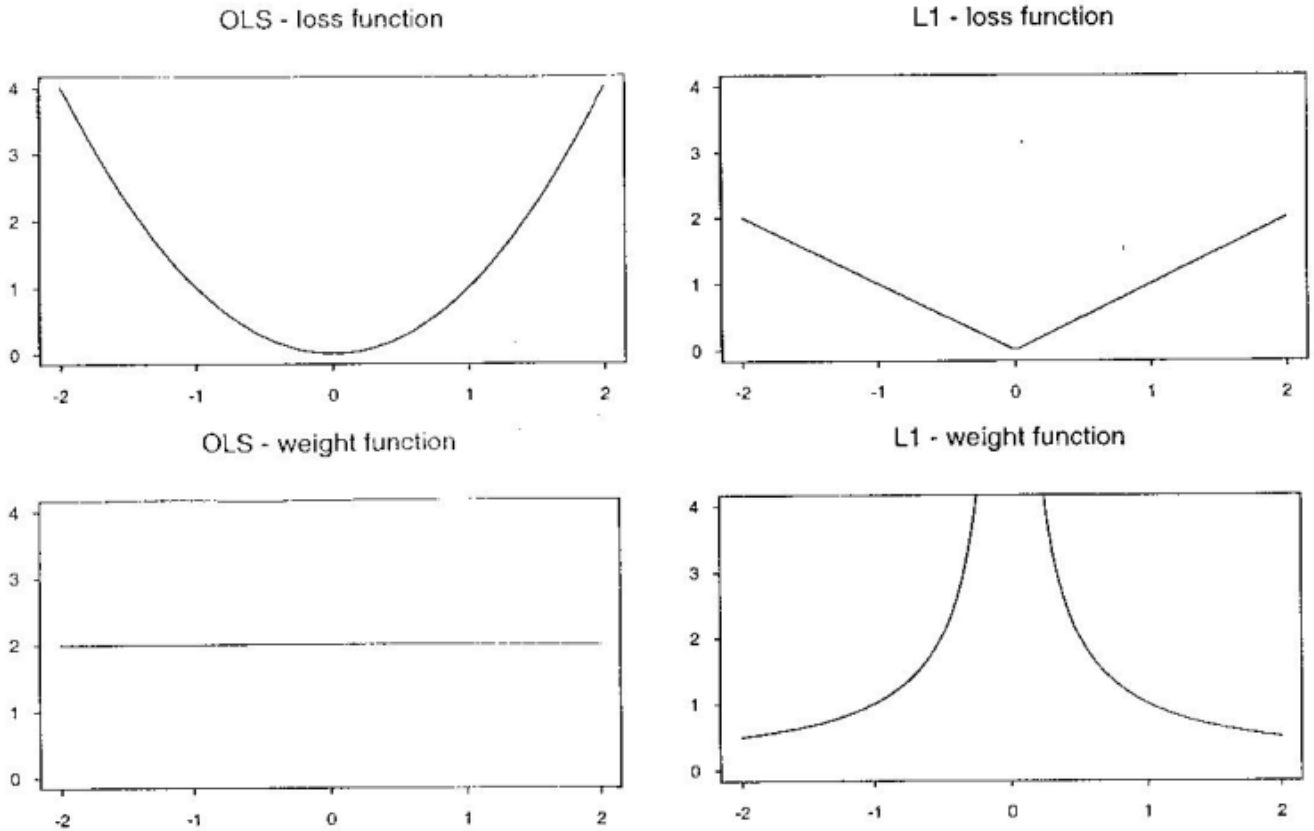
Distance functions are in $1-1$ correspondence with loss functions.

Idea: choose ρ to have desirable properties (reduce/eliminate impact of outliers) — same as choosing weight function.

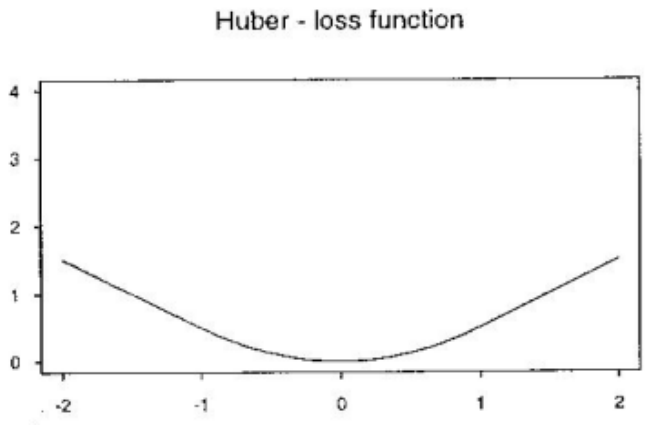
Some typical examples are:

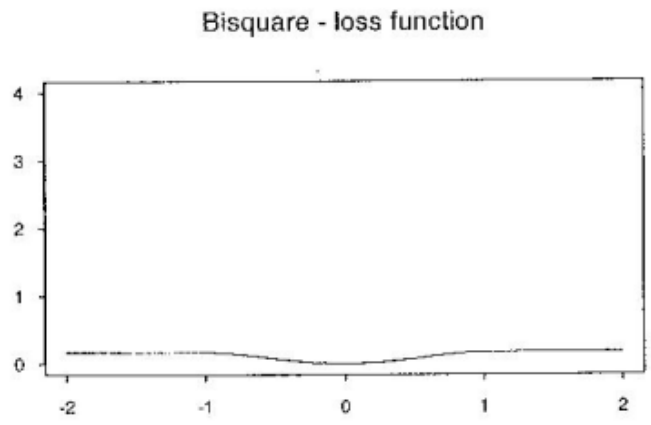
- $\rho(z) = z^2 \qquad \Leftrightarrow w(u) = 2$
- $\rho(z) = |z| \qquad \Leftrightarrow w(u) = 1/|u|$
- Huber: $\rho(z) = \begin{cases} z^2, \text{ if } |z| \leq H \\ 2H|z| - H^2, \text{ otherwise} \end{cases}$
- Bisquare: $\rho(z) = \begin{cases} \frac{1}{6}B^2 \left[1 - \left\{1 - (z/B)^2\right\}^3\right], & |z| \leq B, \\ \frac{1}{6}B^2, \text{ otherwise.} \end{cases}$

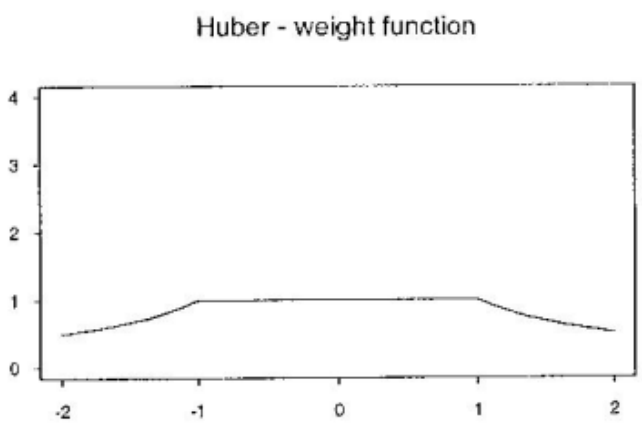
Examples of Distance Functions and Weight Functions

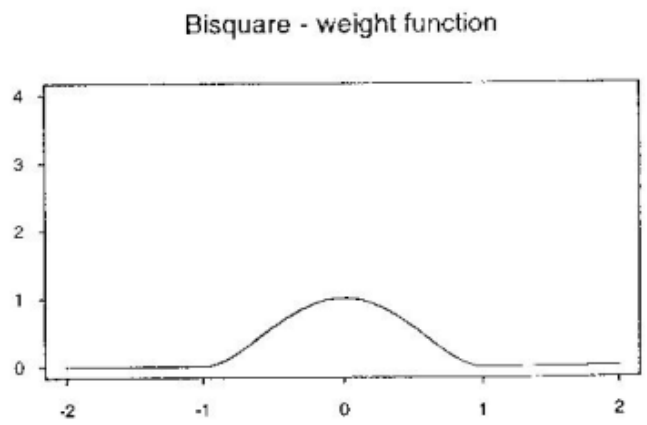


Examples of Distance Functions and Weight Functions









- Explicit expression for LSE
- M-Estimation: non-linear optimisation problem — use iterative approach
- Iteratively re-weighted least squares:

1. Obtain initial estimate $\hat{\beta}^{(0)}$
2. Form "standardised" residuals[3] $u_i^{(0)} = (y_i - x_i^\top \hat{\beta}^{(0)})/\text{MAD}(y_i - x_i^\top \hat{\beta}^{(0)})$
3. Obtain $w_i^{(0)} = w(u_i^{(0)})$ for the chosen weight function $w(\cdot)$
4. Perform weighted least squares with $V^{(0)} = \text{diag}\{w_1^{(0)}, \ldots, w_n^{(0)}\}$
5. Obtain updated estimate $\hat{\beta}^{(1)}$
6. Iterate until convergence (?)

[3] Recall that we want to have a robust notion of scale, to estimate σ

- Obtained M-Estimator as the solution to the system

$$X^\top \boldsymbol{\psi}(\gamma) = 0$$

instead of $X^\top (y - X\gamma) = 0$. Here we defined

$$\boldsymbol{\psi}(\gamma) = \left(\psi\left(\frac{y_1 - x_1^\top \gamma}{\sigma} \right), \ldots, \psi\left(\frac{y_n - x_n^\top \gamma}{\sigma} \right) \right)^\top$$

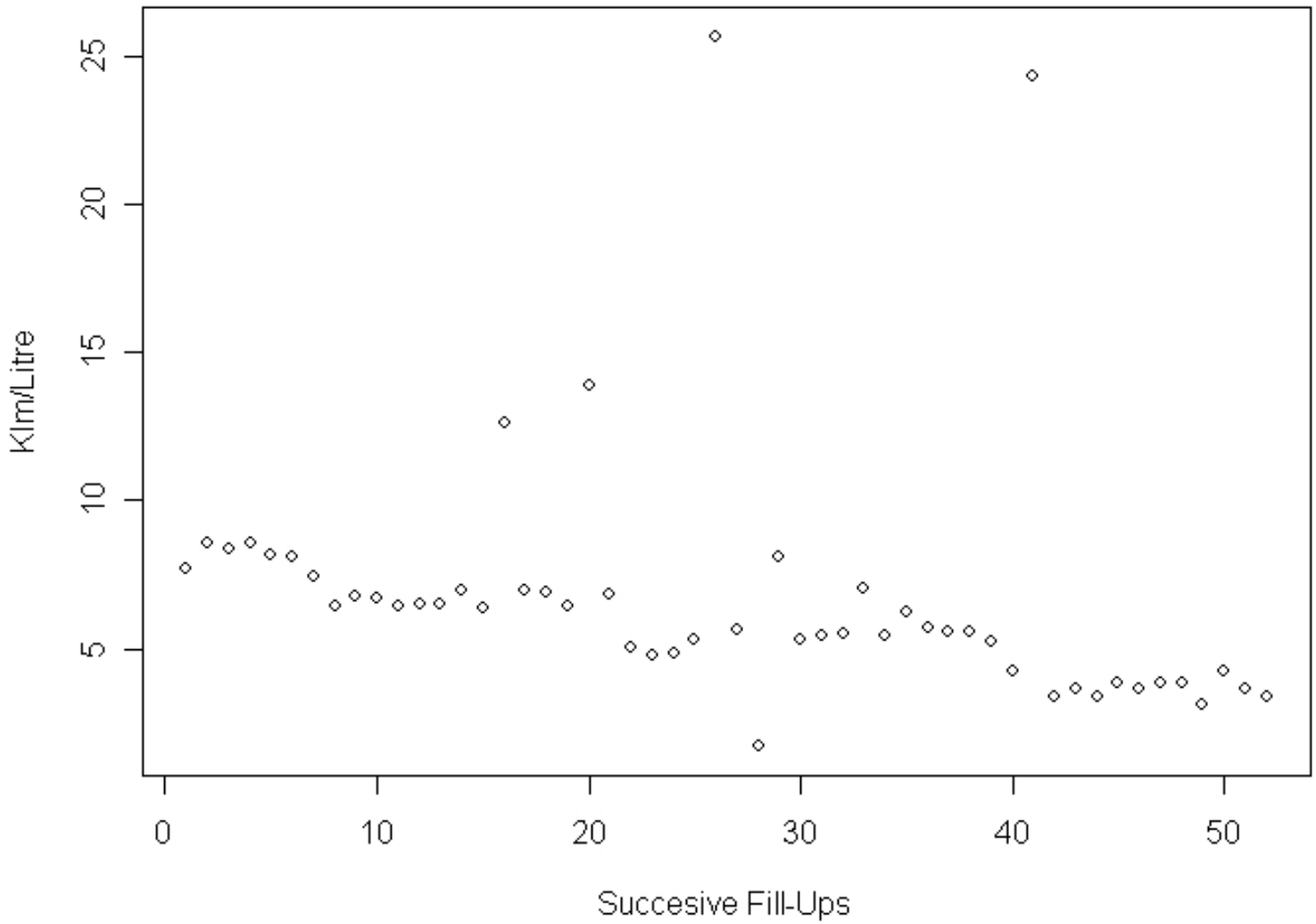
- If these estimating equations are unbiased, i.e.,

$$\mathbb{E}_\beta \left[X^\top \Psi(\beta) \right] = 0, \quad \forall \beta \in \mathbb{R}^p,$$

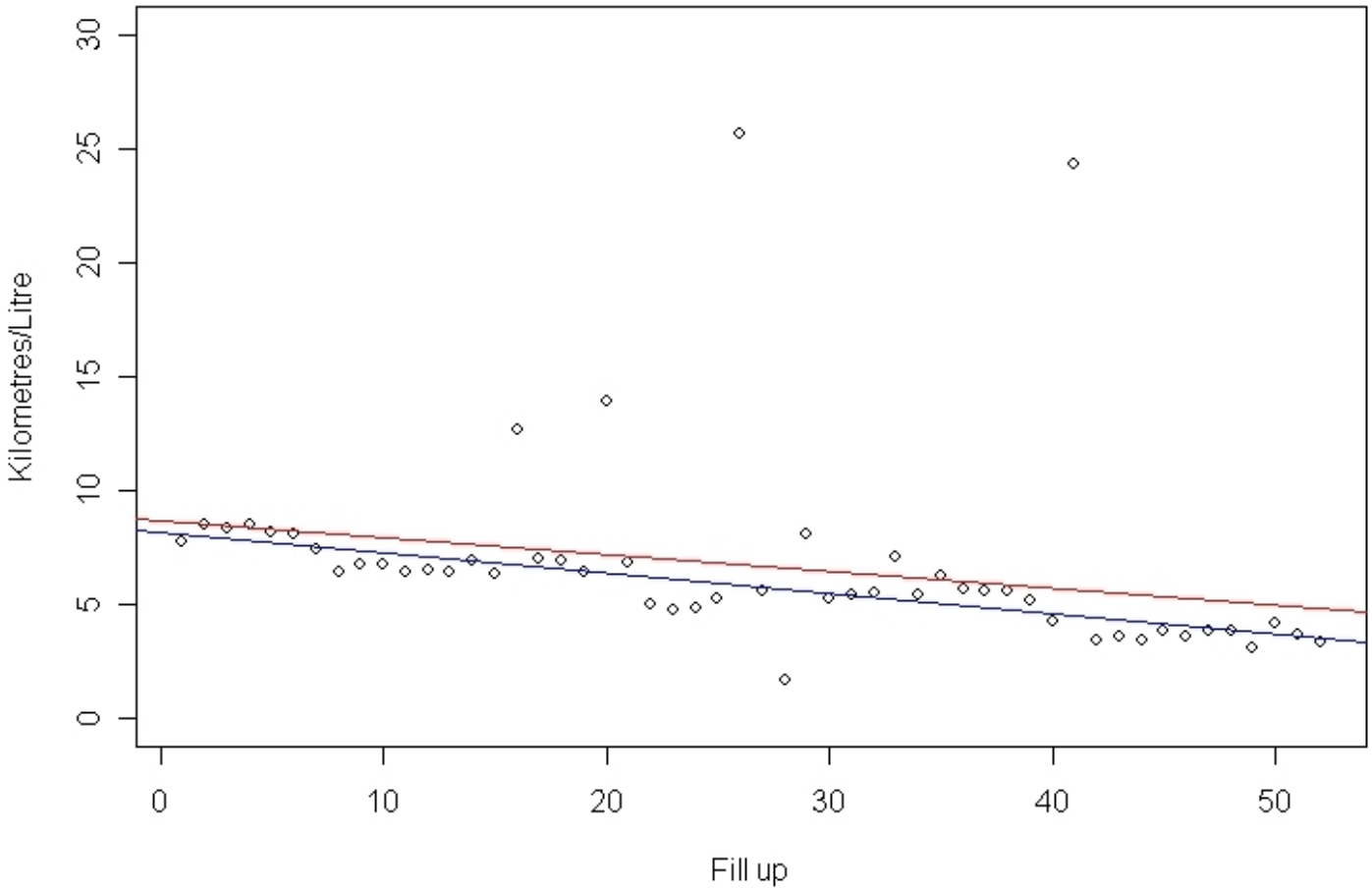
then under mild regularity conditions, as $n \to \infty$, we can show that

$$\hat{\beta}_n \overset{d}{\approx} \mathcal{N}_p \left(\beta, \left\{ \mathbb{E}[X^\top \nabla \boldsymbol{\psi}] \right\}^{-1} X^\top \mathbb{E}[\boldsymbol{\psi}\boldsymbol{\psi}^\top] X \left\{ \mathbb{E}[\nabla \boldsymbol{\psi}^\top X] \right\}^{-1} \right).$$

Example: Professor's Van

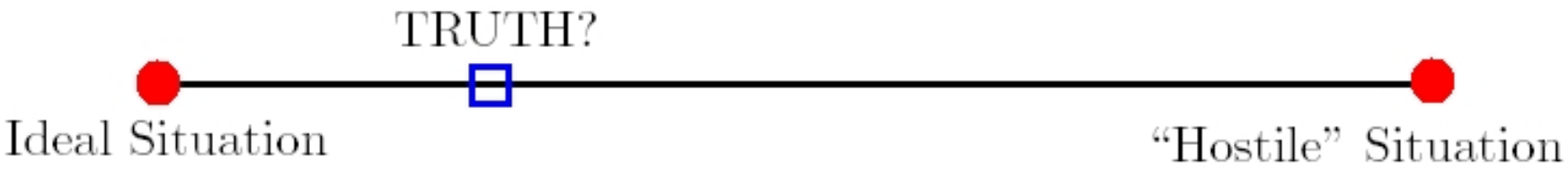


Example: Professor's Van



$\hat{\beta} = -0.07$ (with $p = 0.06$) while $\tilde{\beta} = -0.09$ (with $p \simeq 0$)

Remember our picture:



- ARE measures quality of one estimator of $\theta_{p\times 1}$ relative to another, often the MLE $\hat{\theta}$, for which $\text{var}(\hat{\theta}) = I(\theta)^{-1}$, for large sample size.
- Generally ARE of $\tilde{\theta}$ relative to $\hat{\theta}$ is less than 1 (100%): low ARE is bad, high ARE is good.
- ARE of $\tilde{\theta}$ relative to $\hat{\theta}$ is

$$\left\{ \frac{|\text{var}(\hat{\theta})|}{|\text{var}(\tilde{\theta})|} \right\}^{1/p} \quad (\times 100\%).$$

- ARE of $\tilde{\theta}_r$ relative to $\hat{\theta}_r$ is

$$\frac{\text{var}(\hat{\theta}_r)}{\text{var}(\tilde{\theta}_r)} \quad (\times 100\%).$$

- Linear model $y = X\beta + \varepsilon$, with $\varepsilon_j \overset{iid}{\sim} g(\cdot)$; assume $\text{var}(\varepsilon_j) = \sigma^2 < \infty$ is known.
- Assume MLE is regular, with

$$i_g = \int -\frac{\partial^2 \log g(u)}{\partial u^2} g(u)\, du = \int \left\{ \frac{\partial \log g(u)}{\partial u} \right\}^2 g(u)\, du.$$

- ARE of LSE of β relative to MLE of β is

$$\frac{1}{\sigma^2 i_g}$$

Examples:

- ARE at $g(\cdot)$ Gaussian: 1
- ARE at $g(\cdot)$ Laplace: $1/2$
- ARE of Huber at $g(\cdot)$ Gaussian is 95% with $H = 1.345$

A simple and useful strategy is to perform one's analysis both robustly and by standard methods and to compare the results. If the differences are minor, either set may be presented. If the differences are not minor, one must perforce consider why not, and the robust analysis is already at hand to guide the next steps.

- Perform analysis both ways and compare results.
- Plot weights to see which observations were downweighted.
- Try to understand why.

Nonlinear and Nonparametric Models

Recall most general version of regression given in Week 1:

$$Y_i \mid x_i^\top \overset{ind}{\sim} \text{Dist}\{g(x_i^\top)\}, \quad i = 1, \ldots, n.$$

So far we have investigated what happens when

$$\begin{cases} g(x^\top) = x^\top \beta, & \beta \in \mathbb{R}^p, \\ \text{Dist} = \mathcal{N}(x^\top \beta, \sigma^2). \end{cases}$$

We now consider a more general situation:

$$Y_i \mid x_i^\top \overset{ind}{\sim} \mathcal{N}\{\eta(x_i^\top; \beta), \sigma^2\}, \quad i = 1, \ldots, n,$$

where $\eta(x_i^\top; \beta)$

- is a KNOWN function,
- that depends on a parameter $\beta \in \mathbb{R}^p$,
- but is not linear in β.

- Decennial population data from US, for 1790–1990.
- y is population in millions, x is time.

Regression model:

$$Y_i = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 x_i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \; i = 1, \ldots, n.$$

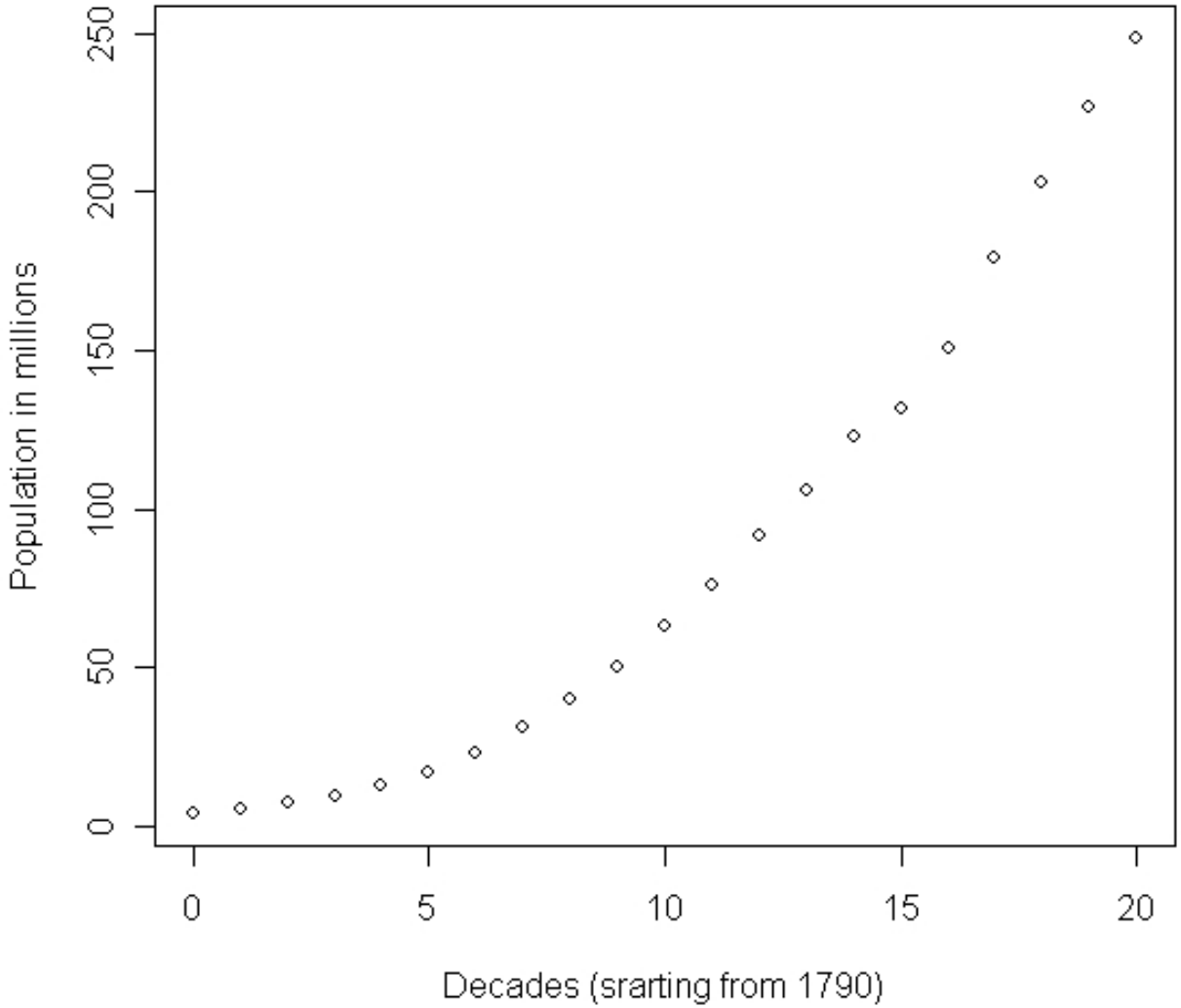
- Here

$$\eta(x; \beta) = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 x)}.$$

- Distribution remains Gaussian.
- Cannot transform into a linear regression problem.
- Coefficient interpretation different than in a linear model.
- Related to the differential equation

$$\frac{d}{dx}\eta(x) = C \times \eta(x)\{1 - \eta(x)\}.$$

Example: Logistic Growth



- Still assume independent random variables $Y_1, \ldots, Y_n$, with observed values $y_1, \ldots, y_n$, and explanatories $x_1, \ldots, x_n$.
- Distribution still Gaussian.

Introduce notation:

- $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,
- $\eta(\beta) = (\eta_1(\beta), \ldots, \eta_n(\beta))^\top = (\eta(x_1^\top, \beta), \ldots, \eta(x_n^\top, \beta))^\top$, i.e.,

$$\eta(\beta) : \mathbb{R}^p \to \mathbb{R}^n \qquad \beta \in \mathbb{R}^p \mapsto \eta(\beta) \in \mathbb{R}^n$$

- Therefore $\eta(\beta)$ is a vector-valued function.
- Analogy with linear case: $\eta(\beta)$ plays the role of $X\beta$ but is no longer linear in β.

Model now is:

$$\underset{n\times 1}{y} = \underbrace{\eta(\beta)}_{n\times 1} + \underset{n\times 1}{\varepsilon}, \qquad \beta \in \mathbb{R}^p, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I).$$

Since $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, have

$$y \sim \mathcal{N}\{\eta(\beta), \sigma^2\},$$

so likelihood and loglikelihood are

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \eta(\beta))^\top (y - \eta(\beta))\right\},$$

$$\ell(\beta, \sigma^2) = -\frac{1}{2}\left\{n \log 2\pi + n \log \sigma^2 + \frac{1}{\sigma^2}(y - \eta(\beta))^\top (y - \eta(\beta))\right\}.$$

. . . exactly as in linear case, but with $\eta(\beta)$ replacing $X\beta$. Hence, suggests *least squares estimators*,

$$\begin{cases} \hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - \eta(\beta)\|^2 & \text{(assuming identifiability)}, \\ \hat{\sigma}^2 = \frac{1}{n}\|y - \eta(\hat{\beta})\|^2. \end{cases}$$

Main problem is *non-linearity* — cannot obtain closed form solution in general.

↪ Idea: linearise locally, assuming that η is sufficiently smooth.

First-order Taylor expansion: approximate as

$$\underset{n\times 1}{\eta(\beta)} \simeq \underset{n\times 1}{\eta(\beta^{(0)})} + \underbrace{[\nabla_\beta \eta]_{\beta=\beta^{(0)}}}_{n\times p} \underbrace{(\beta - \beta^{(0)})}_{p\times 1}$$

where β is sufficiently close to $\beta^{(0)}$.

- We dropped higher order terms by appealing to smoothness of η (smoothness $\iff$ "close to zero" higher derivatives).

Linearised representation suggests Newton–Raphson iteration:

- Suppose an initial estimate $\beta^{(0)}$ is available ($||\beta^{(0)} - \hat{\beta}|| < \epsilon$).
- Let $D^{(0)} = [\nabla_\beta \eta]_{\beta=\beta^{(0)}}$ and $\beta = u^{(0)} + \beta^{(0)}$.
- Taylor expansion yields

$$y - \eta(\beta^{(0)}) \approx D^{(0)} \underbrace{(\beta - \beta^{(0)})}_{u^{(0)}} + \varepsilon.$$

To get β we need $u^{(0)}$. Consider the following iteration:

1. Initialise with $\beta^{(0)}$.
2. Let $u^{(1)} = \underset{u \in \mathbb{R}^p}{\arg\min} ||y - \eta(\beta^{(0)}) - D^{(0)} u||^2$

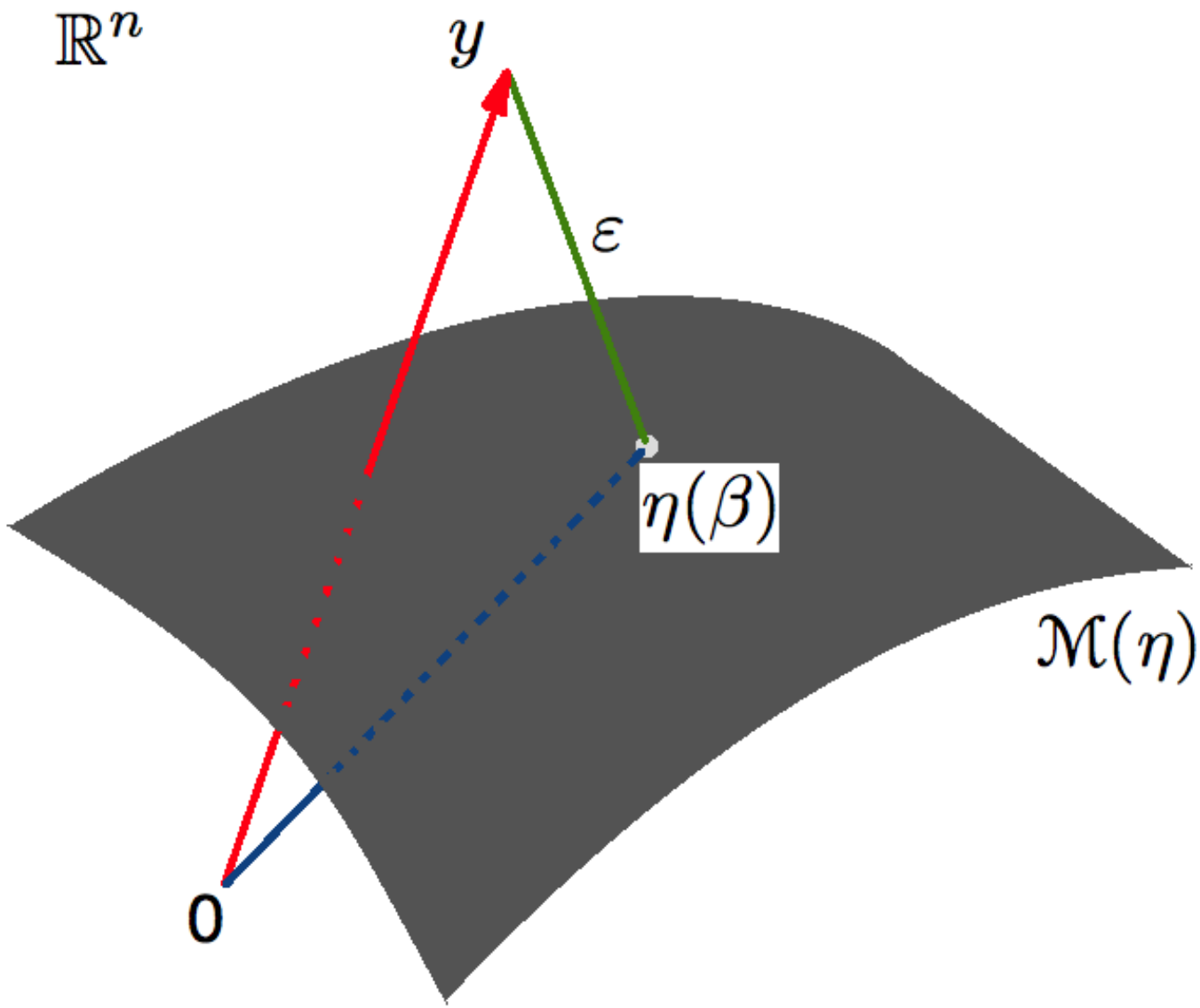
:-) (but this is just a linear least squares problem, with $y^{(0)} = y - \eta(\beta^{(0)})$ and $X^{(0)} = D^{(0)}$!)

3. Thus set $u^{(1)} = ([D^{(0)}]^\top D^{(0)})^{-1} [D^{(0)}]^\top \{y - \eta(\beta^{(0)})\}$.
4. Let $\beta^{(1)} = \beta^{(0)} + u^{(1)}$ and iterate until convergence criterion satisfied. Return last $\beta^{(k)}$ as $\hat{\beta}$.

As β ranges over $\mathbb{R}^p$, $\eta(\beta)$ traces a p-dimensional differentiable manifold (smooth surface) in $\mathbb{R}^n$,

$$\mathcal{M}(\eta) = \{\eta(\beta) : \beta \in \mathbb{R}^p\}.$$

- β provides the intrinsic coordinates on that manifold.
- y is obtained by selecting a point $\eta(\beta)$ on the manifold, and adding a mean zero Gaussian vector ε.
- Regression asks to find the coordinates of the point on the manifold that generated y.
- Would like to project y on the manifold, but do not have a closed form expression!


$\mathbb{R}^n$
y
ε
$\eta(\beta)$
$\mathcal{M}(\eta)$
0

Newton–Raphson algorithm is interpretable via differential geometry:

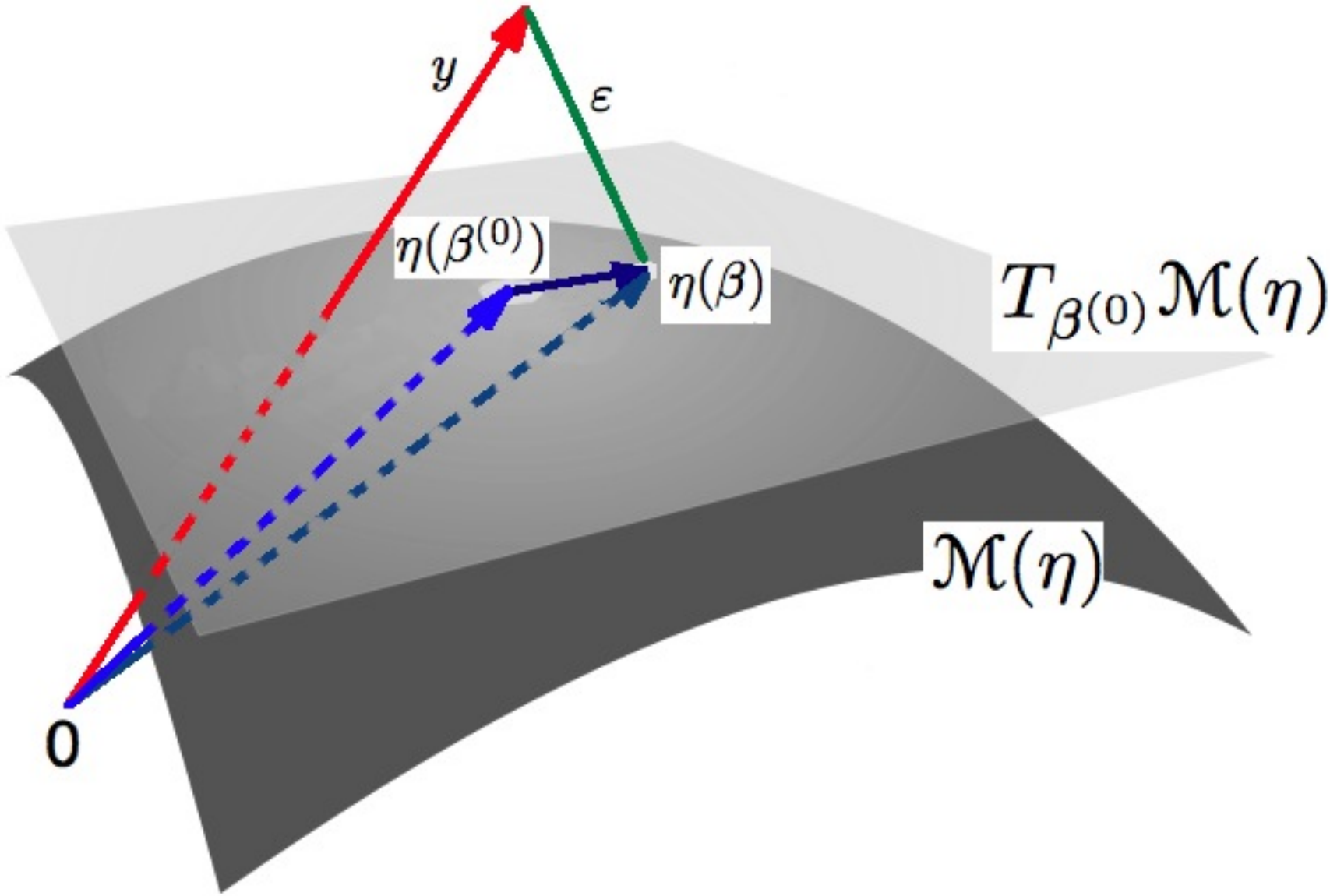
- The p-dimensional tangent plane at a point $\eta(\beta^{(0)}) \in \mathcal{M}(\eta)$ is spanned by $\eta(\beta^{(0)}) + [\nabla_\beta \eta(\beta)]_{\beta=\beta^{(0)}} u, \qquad u \in \mathbb{R}^p$.
- Hence we may write that

$$T_{\beta^{(0)}} \mathcal{M}(\eta) = \{\eta(\beta^{(0)}) + D^{(0)} u : u \in \mathbb{R}^p\}$$

- In other words, the p columns of $D^{(0)}$, when translated by $\eta(\beta^{(0)})$, form a basis for the tangent plane at $\eta(\beta^{(0)})$.
- Taylor expansion merely says that if β is close to $\beta^{(0)}$, we approximately have $\eta(\beta) - \eta(\beta^{(0)}) \in T_{\beta^{(0)}} \mathcal{M}(\eta)$. This is equivalent to the expression

$$\eta(\beta) - \eta(\beta^{(0)}) \approx \underbrace{[\nabla_\beta \eta]_{\beta=\beta^{(0)}}}_{D^{(0)}} \underbrace{(\beta - \beta^{(0)})}_{u^{(0)}}.$$

- Therefore, $y - \eta(\beta^{(0)}) \approx D^{(0)} u^{(0)} + \varepsilon$ means that $\mathbb{E}[y]$ approximately lies in $T_{\beta^{(0)}} \mathcal{M}(\eta)$.
- Newton–Raphson algorithm $\equiv$ iterated projection on approximating linear subspaces.


y
ε
$\eta(\beta^{(0)})$
$\eta(\beta)$
$T_{\beta^{(0)}}\mathcal{M}(\eta)$
$\mathcal{M}(\eta)$
0

- Summarising, suppose we consider $\eta(\beta^{(0)})$ as the origin of space (i.e., now the tangent space is a subspace).
- Then $y - \eta(\beta^{(0)})$ is approximately the response obtained when adding ε to an element $D^{(0)}(\beta - \beta^{(0)}) \in T_{\beta_{(0)}}\mathcal{M}(\eta)$.
- So, approximately, we have our usual linear problem, and we can use orthogonal projection to solve it.
- Amounts to approximating the manifold $\mathcal{M}(\eta)$ by a plane $T_{\beta_{(0)}}\mathcal{M}(\eta)$ locally around $\eta(\beta^{(0)})$.

Once initial value $\beta^{(0)}$ is updated to $\beta^{(1)}$, use a new tangent plane approximation and repeat the whole procedure.

But how do we obtain our initial $\beta^{(0)}$?

Successful linearisation depends on good initial value.

- Occasionally, can find initial values by inspection in simple problems.
- More generally, it takes some experimentation.
  - E.g., one can try fitting polynomial models to data.
  - Use these to find fitted values at fixed design points.
  - Solve a system of equations to get initial values.

Example: consider the model $y_j = \beta_0 + \beta_1 \exp\{(-x_j/\theta)\} + \varepsilon_j$

1. Fit a polynomial regression to data
2. Find fitted values $\tilde{y}_0, \tilde{y}_1, \tilde{y}_2$ at $x_0, x_0 + \delta, x_0 + 2\delta$.
3. Equate fitted values with model expectation:

$$\tilde{y}_k = \beta_0 + \beta_1 \exp\{-(x_0 + k\delta)/\theta\}, \quad k = 0, 1, 2.$$

4. System yields initial estimate $\theta^{(0)} = \delta / \log\left[(\tilde{y}_0 - \tilde{y}_1)/(\tilde{y}_1 - \tilde{y}_2)\right]$
5. Get initial values for β_0, β_1 by linear regression, once $\theta^{(0)}$ is at hand.

Under smoothness conditions on η, one can in general prove that

$$\boxed{S^{-1}\left\{\nabla_\beta \eta(\hat{\beta})^\top \nabla_\beta \eta(\hat{\beta})\right\}^{1/2}(\hat{\beta}-\beta) \overset{d}{\approx} N_p(0, I_p)}$$

for large n, where $S = (n-p)^{-1}\|e\|^2$. May thus mimic linear case:

$$c^\top \hat{\beta} \overset{d}{\approx} \mathcal{N}_1\left[c^\top \beta, S^2 c^\top \left\{\nabla_\beta \eta(\hat{\beta})^\top \nabla_\beta \eta(\hat{\beta})\right\}^{-1} c\right].$$

So base confidence intervals (and tests) on

$$\frac{c^\top \hat{\beta} - c^\top \beta}{\sqrt{S^2 c^\top \left\{\nabla_\beta \eta(\hat{\beta})^\top \nabla_\beta \eta(\hat{\beta})\right\}^{-1} c}} \overset{d}{\approx} N(0,1),$$

which gives a $(1-\alpha)\times 100\%$ CI:

$$\boxed{c^\top \hat{\beta} \pm z_{\alpha/2}\sqrt{S^2 c^\top \left\{\nabla_\beta \eta(\hat{\beta})^\top \nabla_\beta \eta(\hat{\beta})\right\}^{-1} c}}.$$

Until today we have discussed the following setup:

$$Y_i \mid x_i \overset{ind}{\sim} \text{Dist}[y \mid \theta_i] \to \begin{cases} \theta_i = g(x_i; \beta), \\ \beta \in \mathcal{B} \subset \mathbb{R}^p, \end{cases}$$

with $g(\cdot; \beta)$ known up to β to be estimated from data, e.g.

- $\text{Dist}(\cdot \mid \mu) = \mathcal{N}(\cdot \mid \mu)$ and $\mu = g(x \mid \beta) = x^\top \beta$,
- $\text{Dist}(\cdot \mid \mu) = \mathcal{N}(\cdot \mid \mu)$ and $\mu = g(x \mid \beta) = \eta(x; \beta)$.

Would now like to extend model to a more flexible dependence:

$$Y_i \mid x_i \overset{ind}{\sim} \text{Dist}[y \mid \theta_i] \to \begin{cases} \theta_i = g(x_i), \\ g \in \mathcal{F} \subset L^2(\mathbb{R}^p) \text{ (say)}, \end{cases}$$

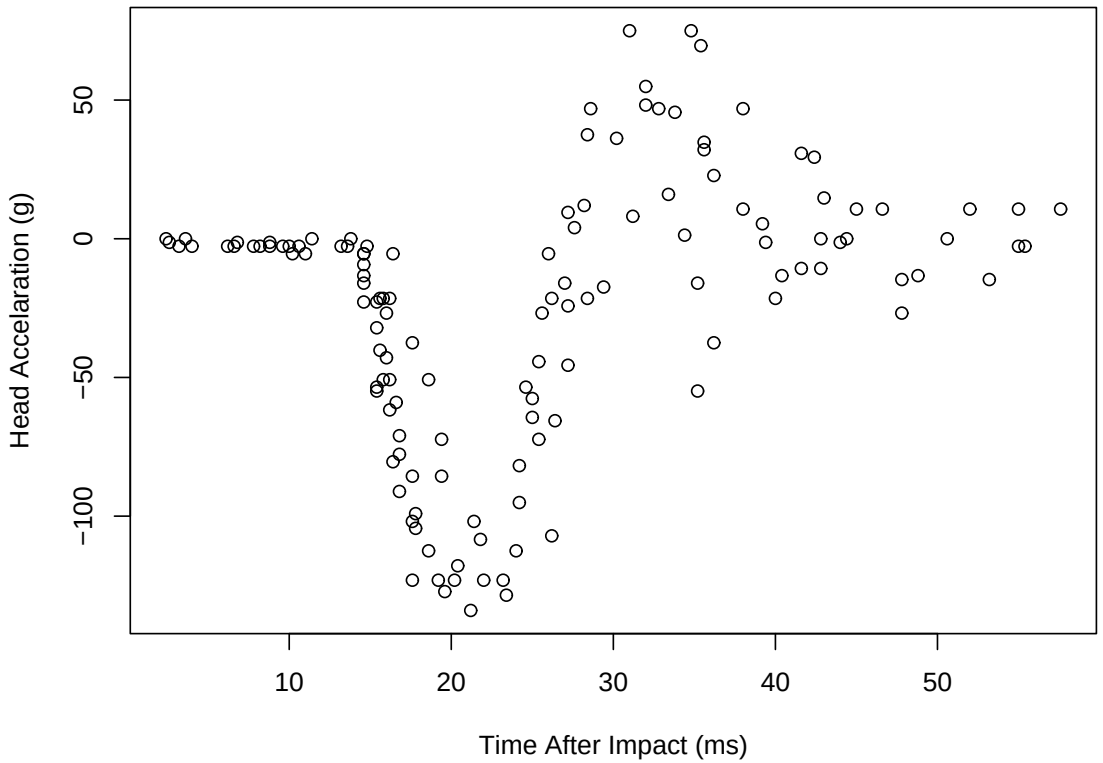
with g unknown, to be estimated given data $\{(y_i, x_i)\}_{i=1}^n$.

- A *nonparametric* problem (parameter ∞-dimensional)!
- How to estimate g in this context?
- $\mathcal{F}$ is usually assumed to be a class of smooth functions (e.g., C^k).

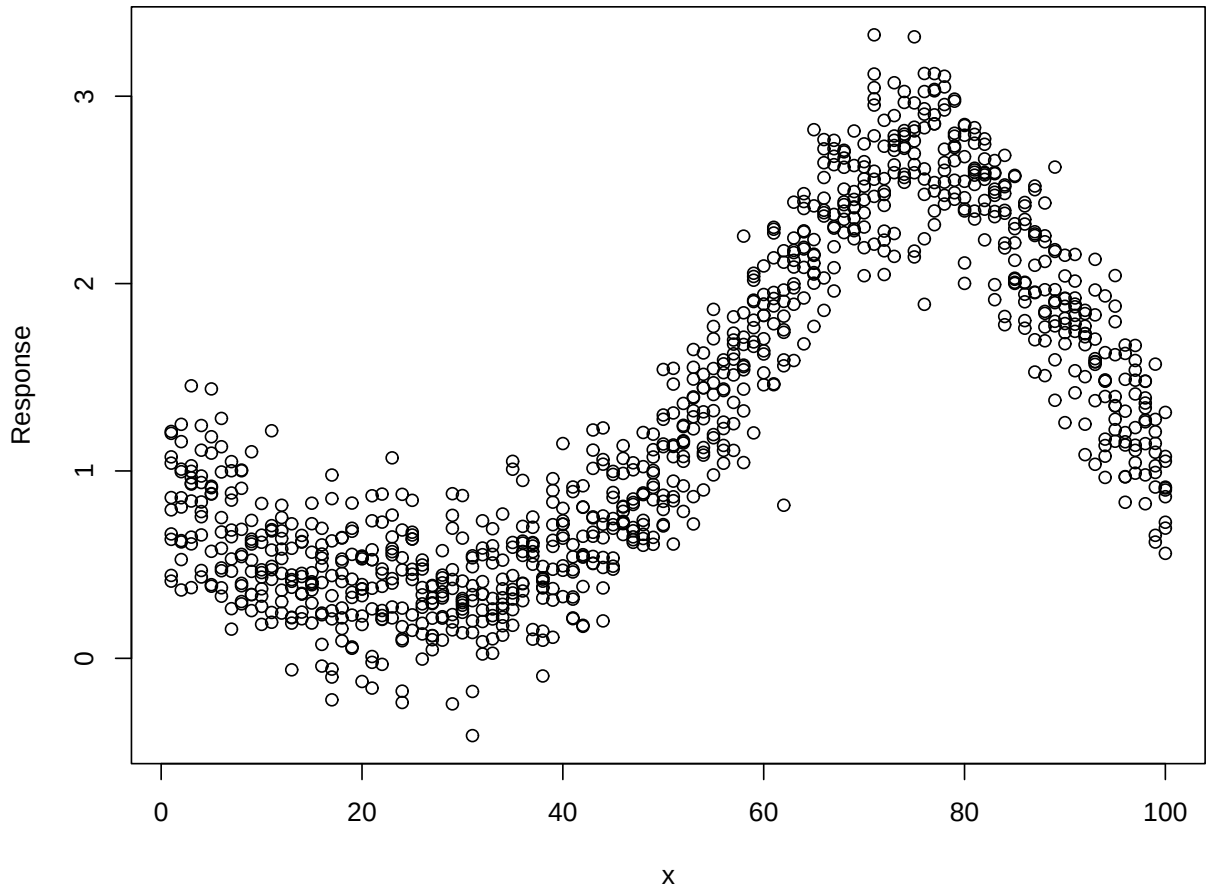
Start from simplest problem:

$$\left.\begin{array}{l}\text{Dist} \equiv \mathcal{N}(\mu, \sigma^2) \\ x_i \in \mathbb{R}\end{array}\right\} \implies Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Figure: Motorcycle Accident Data

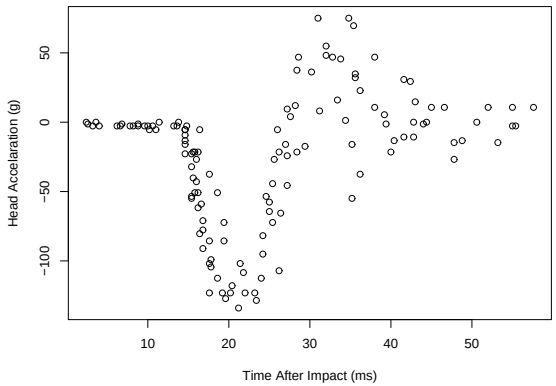


- Ideally: multiple y's at each x_i ($n \to \infty$ and large *covariate classes*):



- Then average y's at each x_i and interpolate ...
- But this is never the case ...

- Usually unique x_i distinct:



- Here is where the smoothness assumption comes in
- Since have unique y at each x_i, need to borrow information from nearby . . .
- . . . use continuity!!! (or even better, *smoothness*)

▶ Recall: A function $g : \mathbb{R} \to \mathbb{R}$ is *continuous* if:

$$\forall\ \epsilon > 0\ \exists\ \delta > 0 :\ |x - x_0| < \delta \implies |g(x) - g(x_0)| < \epsilon.$$

▶ So maybe average y_i's corresponding to x_i's in a δ-neighbourhood of x as $\hat{g}(x)$?

▶ Motivates the use of a kernel smoother . . .

Naive idea: $\hat{g}(x_0)$ should be the average of y_i-values with x_i's "close" to x_0.

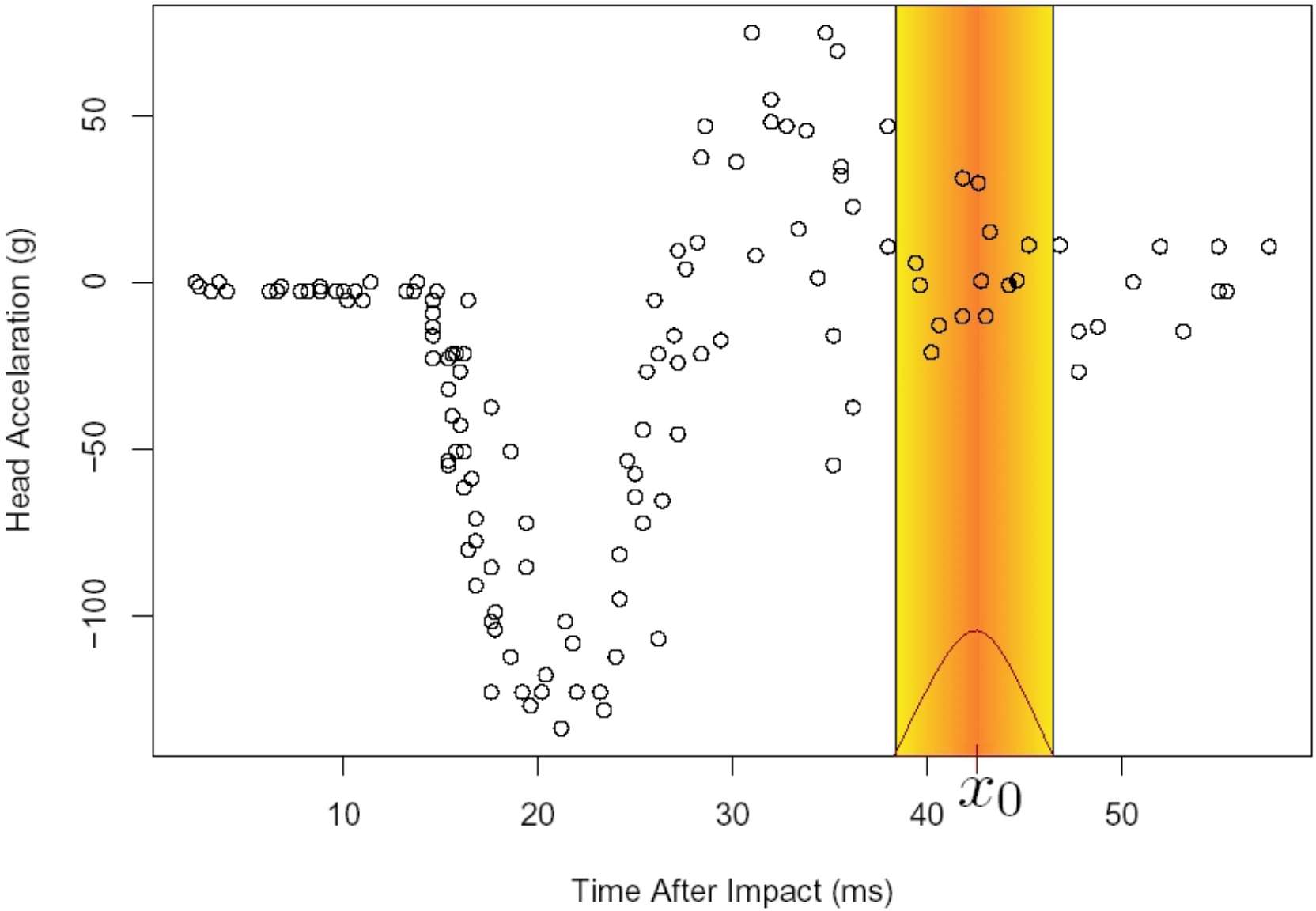
$$\hat{g}(x_0) = \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{|x_i - x_0| \leq \lambda\}} \sum_{i=1}^{n} y_i \mathbf{1}\{|x_i - x_0| \leq \lambda\}.$$

A weighted average! Choose other weights? Kernel estimator:

$$\hat{g}(x_0) = \frac{1}{\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{\lambda}\right)} \sum_{i=1}^{n} y_i K\left(\frac{x_i - x_0}{\lambda}\right).$$

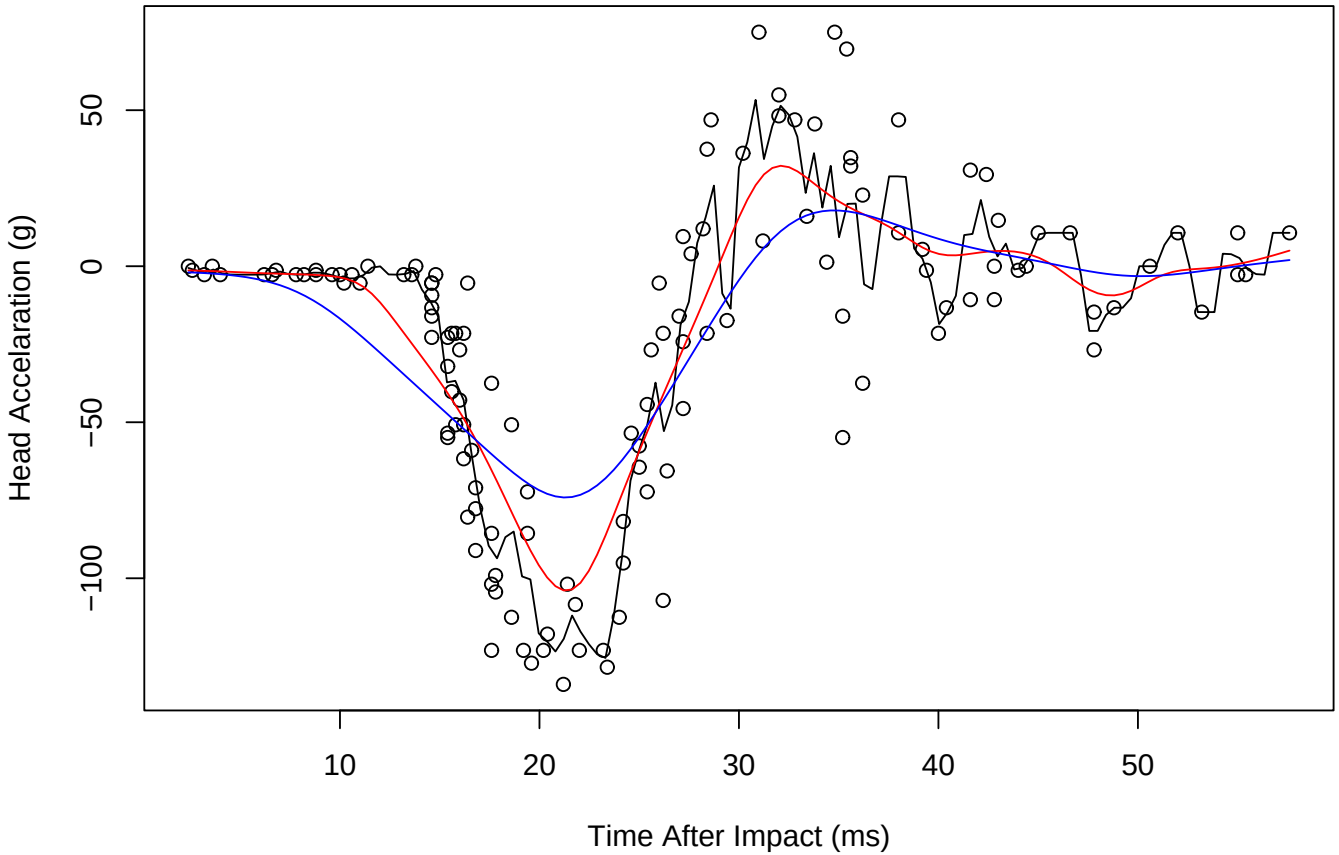
- K is a weight function (kernel), e.g. a pdf
  - ↪ Usually symmetric, non-negative, decreasing away from zero
- λ is the bandwidth parameter
  - ↪ small λ gives local behaviour, large λ gives global behaviour
- Choice of K not so important, choice of λ very important!
- The resulting fitted values are linear in the responses, i.e., $\hat{y} = S_\lambda y$, where the smoothing matrix S_λ depends on $x_1, \ldots, x_n$, K and λ. Analogous to a projection matrix in linear regression, but S_λ is NOT a projection.

Visualising a Kernel at Work



Motorcycle Data Kernel Smooth

```
> plot(time,accel,xlab="Time After Impact (ms)",ylab="Head Accelaration (g)")
> lines(ksmooth(time,accel,kernel="normal",bandwidth=0.7))
> lines(ksmooth(time,accel,kernel="normal",bandwidth=5),col="red")
> lines(ksmooth(time,accel,kernel="normal",bandwidth=10),col="blue")
```



Find $g \in C^2$ that minimises

$$\underbrace{\sum_{i=1}^{n}\{y_i - g(x_i)\}^2}_{\text{Fit Penalty}} + \underbrace{\lambda \int_I \{g''(t)\}^2 dt}_{\text{Roughness Penalty}}$$

- This is a Gaussian likelihood with a roughness penalty
  - ↪ If use only likelihood, any interpolating function is an MLE!
- λ to balance fidelity to the data and smoothness of the estimated h.

Remarkably, problem has unique explicit solution!

↪ Natural Cubic Spline with knots at $\{x_i\}_{i=1}^n$:

- piecewise polynomials of degree 3,
- with pieces defined at the knots,
- with two continuous derivatives at the knots,
- and linear outside the data boundary.

Can represent splines via a basis B_j, as

$$s(t) = \sum_{j=1}^{n} \gamma_j B_j(t).$$

For example, one basis (the natural basis) is

$$\begin{aligned} B_1(t) &= 1 \\ B_2(t) &= t \\ B_{m+2}(t) &= \delta_m(t) - \delta_{n-1}(t), \qquad m = 1, \ldots, n-2 \\ \delta_k(t) &= \frac{(t - x_k)_+^3 - (t - x_n)_+^3}{t_n - t_k}, \qquad k = 1, \ldots, n-1 \end{aligned}$$

where x_m are the knot locations and

$$(\cdot)_+ = \max\{\cdot, 0\}$$

is the positive part of any function.

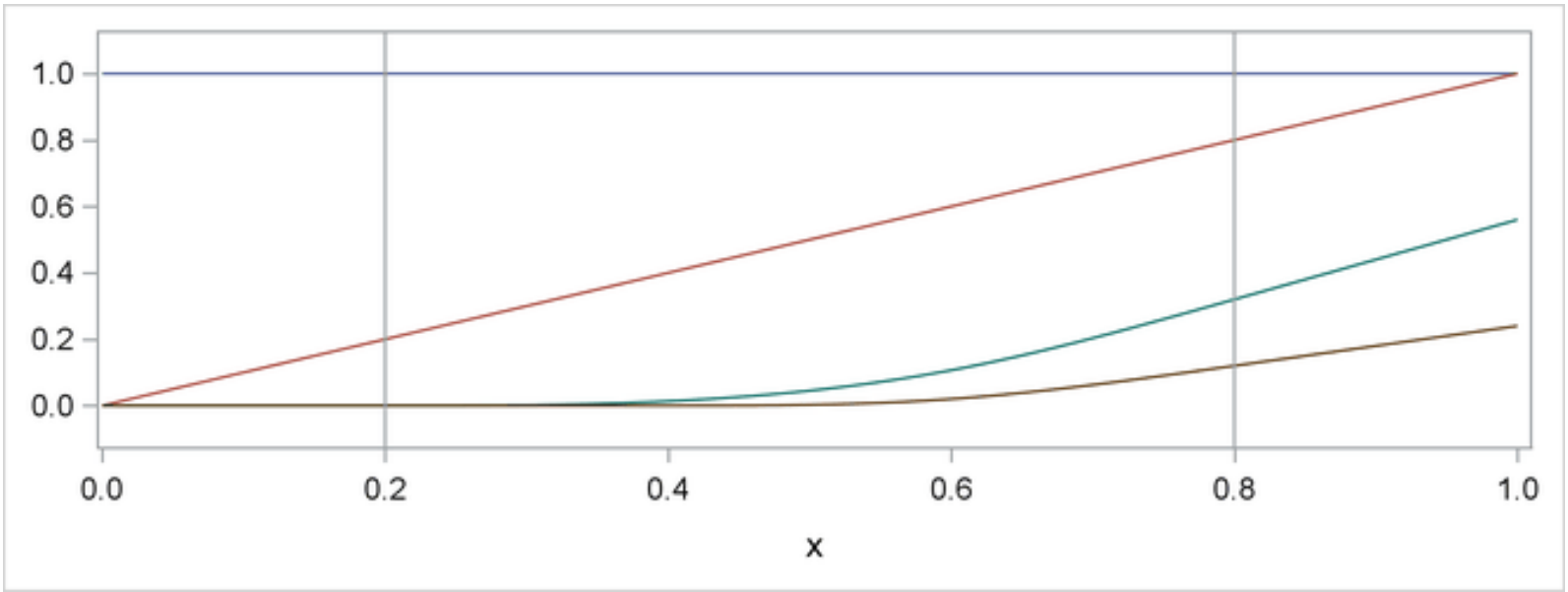


Figure: The $n = 4$ natural spline basis functions for knots at $x_1 = 0.2$, $x_2 = 0.4$, $x_3 = 0.6$ and $x_4 = 0.8$

We wish to find a basis for natural cubic splines with knot locations $\{x_i\}_{i=1}^n$

- Observe that any piecewise polynomial $PP_3(t)$ of order 3 with 2 cts derivatives at the knots can be expanded in the truncated power series basis

$$PP_3(t) = \sum_{j=0}^{3} \phi_j t^j + \sum_{i=1}^{n} \theta_k (t - x_i)_+^3$$

- The $n+4$ coefficients $\{\phi_j\}_{j=0}^3 \cup \{\theta_i\}_{i=1}^n$ must satisfy constraints to ensure linearity beyond boundary knots:
  - $\phi_2 = 0$ & $\phi_3 = 0$
  - $\sum_{i=1}^n \theta_i = 0$
  - $\sum_{i=1}^n \theta_i x_i = 0$

- Can then use relations re-express basis in form on previous slide, with only n (rather than $n+4$) basis functions, and unconstrained coefficients.

Letting $\gamma = (\gamma_1, \ldots, \gamma_n)^\top$,

$$g(t) = \sum_{i=1}^{n} \gamma_i B_i(t), \quad B = \{B_{ij}\} = \{B_j(x_i)\}, \quad \Omega_{ij} = \int B_i''(t) B_j''(t)\, dt,$$

our penalised likelihood

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 \quad + \lambda \int_I \{h''(t)\}^2\, dt$$

becomes

$$\left\{(y - B\gamma)^\top (y - B\gamma) + \lambda \gamma^\top \Omega \gamma\right\}.$$

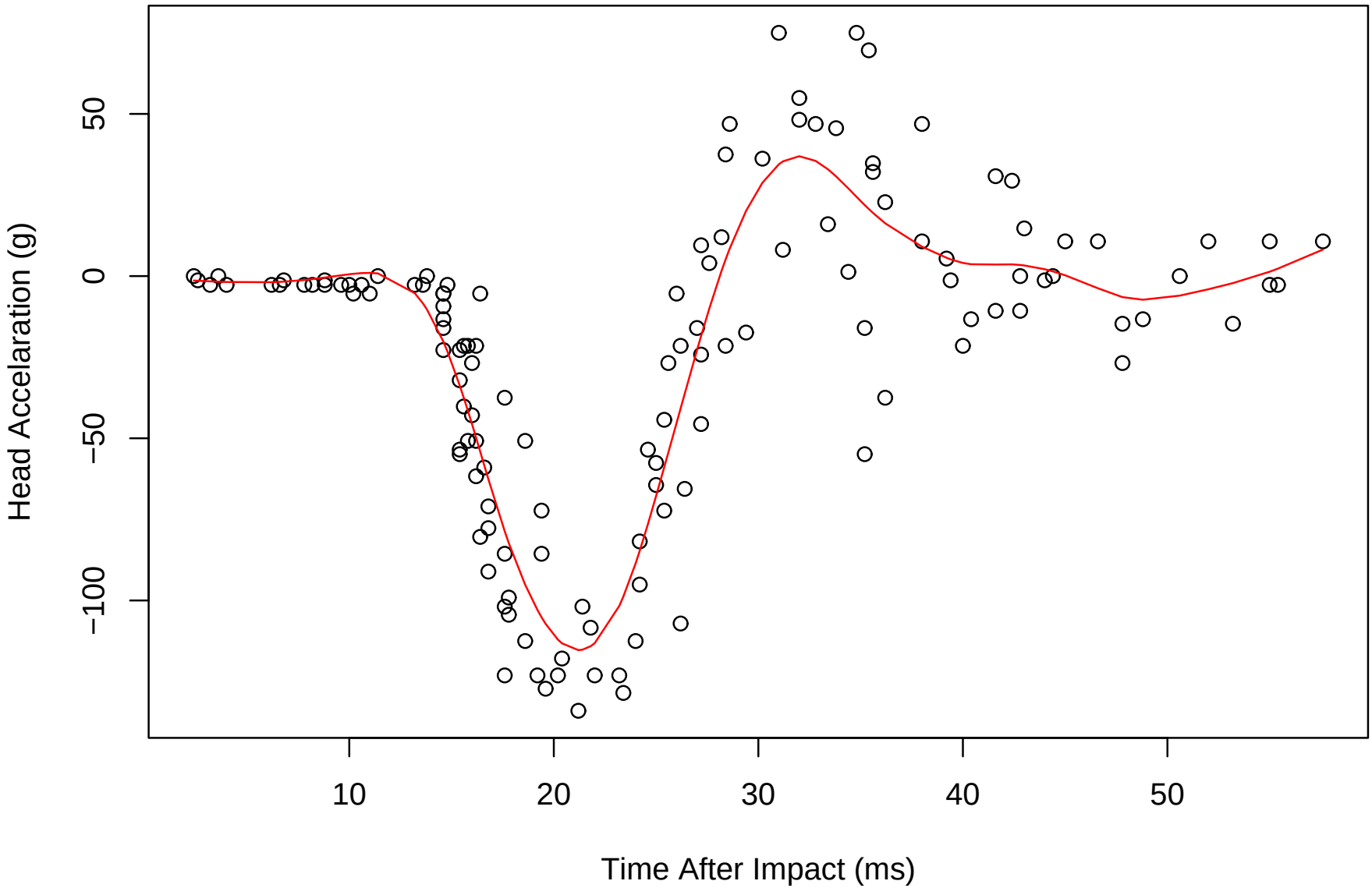
Differentiating and equating with zero yields

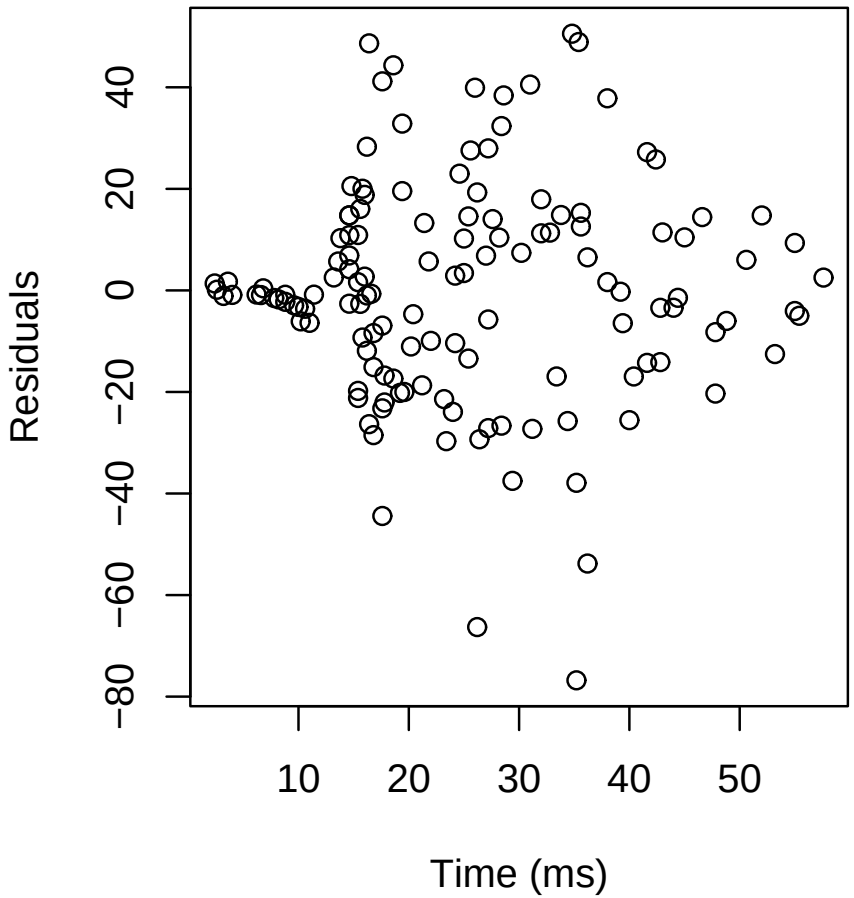
$$(B^\top B + \lambda\Omega)\hat{\gamma} = B^\top y \implies \hat{\gamma} = (B^\top B + \lambda\Omega)^{-1} B^\top y.$$

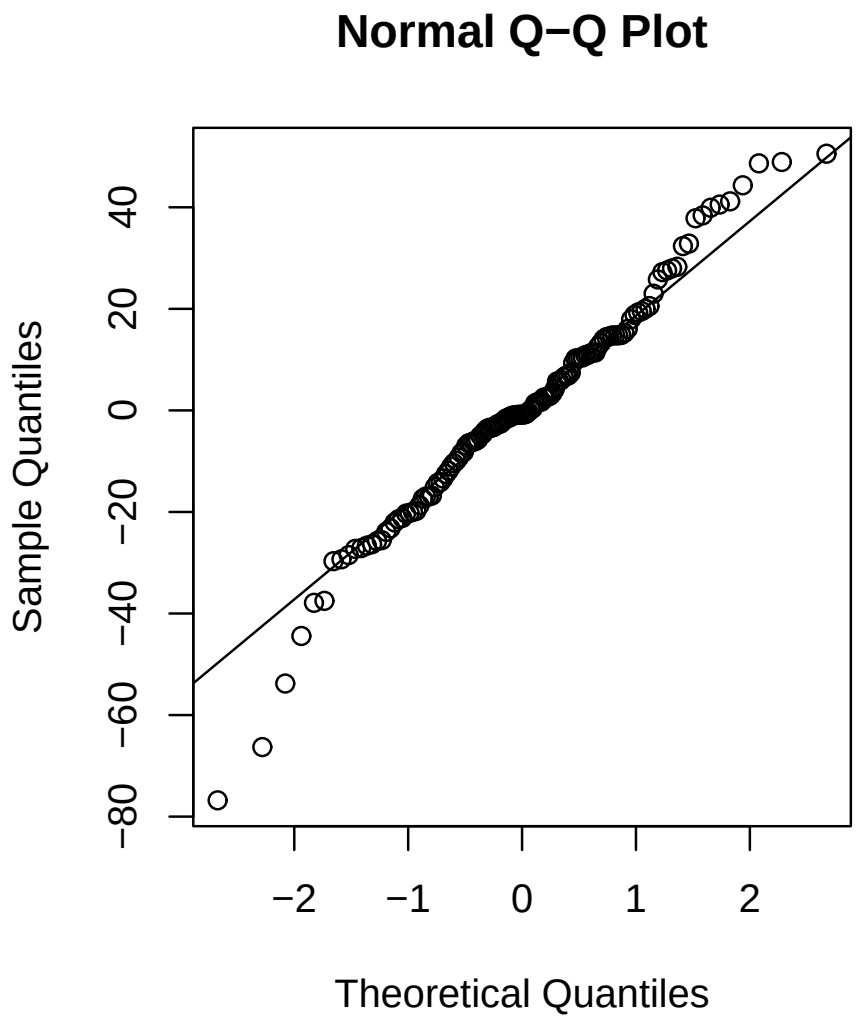
(assuming $B^\top B$ is invertible, which will be shown later to be the case)

- The *smoothing matrix* is $S_\lambda = B(B^\top B + \lambda\Omega)^{-1} B^\top$.
- The natural cubic spline fit is approximately a kernel smoother.

```
lines(smooth.spline(time,accel),col="red")
```




Residuals
−80
−60
−40
−20
0
20
40
10
20
30
40
50
Time (ms)


Normal Q−Q Plot
Sample Quantiles
−80
−60
−40
−20
0
20
40
−2
−1
0
1
2
Theoretical Quantiles

- Least squares estimation: $y = X_{n\times p}\beta + \varepsilon$, we have $\hat{y} = Hy$, with $\text{trace}(H) = p$, in terms of the projection matrix $H = X(X^\top X)^{-1}X^\top$. Here

$$\hat{y} = \underbrace{B(B^\top B + \lambda\Omega)^{-1}B^\top}_{S_\lambda} y.$$

- Idea: define *equivalent degrees of freedom* of smoother

$$\begin{aligned}\text{tr}(S_\lambda) &= \text{tr}\{B(B^\top B + \lambda\Omega)^{-1}B^\top\} = \text{tr}\{B^\top B(B^\top B + \lambda\Omega)^{-1}\} \\ &= \text{tr}\{(B^\top B)^{1/2}(B^\top B + \lambda\Omega)^{-1}(B^\top B)^{1/2}\} \\ &= \text{tr}\{(I + (B^\top B)^{-1/2}\Omega(B^\top B)^{-1/2})^{-1}\} = \sum_{j=1}^{n}\frac{1}{1+\lambda\eta_j}\end{aligned}$$

  where η_j are eigenvalues of $K = (B^\top B)^{-1/2}\Omega(B^\top B)^{-1/2}$.
  (assuming $B^\top B$ is invertible, which will be shown later to be the case)
- Hence $\text{trace}(S_\lambda)$ is monotone decreasing in λ, with $\text{trace}(S_\lambda) \to 2$ as $\lambda \to \infty$ (K will have twos zero eigenvalues) and $\text{trace}(S_\lambda) \to n$ as $\lambda \to 0$.
  Note 1–1 map $\lambda \leftrightarrow \text{trace}(S_\lambda) =$ df, so usually determine roughness using df (interpretation easier).
- Eigenvalues of S_λ lie in $(0, 1)$: it is a smoothing, NOT a projection, matrix.

Focus on the fit for the given grid $x_1, \ldots, x_n$:

$$\hat{\mathbf{g}} = (\hat{g}(x_1), \ldots, \hat{g}(x_n)), \quad \mathbf{g} = (g(x_1), \ldots, g(x_n))$$

Consider the mean squared error:

$$\mathbb{E}(\|\mathbf{g} - \hat{\mathbf{g}}\|^2) = \underbrace{\mathbb{E}\{\|\mathbb{E}(\hat{\mathbf{g}}) - \hat{\mathbf{g}}\|^2\}}_{\text{variance}} + \underbrace{\|\mathbf{g} - \mathbb{E}(\hat{\mathbf{g}})\|^2}_{\text{bias}^2}.$$

When estimator potentially biased, need to worry about both!

In the case of a linear smoother, for which $\hat{\mathbf{g}} = S_\lambda y$, we find that

$$\mathbb{E}(\|\mathbf{g} - \hat{\mathbf{g}}\|^2) = \frac{\text{trace}(S_\lambda S_\lambda^\top)}{n}\sigma^2 + \frac{(\mathbf{g} - S_\lambda \mathbf{g})^\top (\mathbf{g} - S_\lambda \mathbf{g})}{n},$$

so

- $\lambda \uparrow \implies$ variance $\downarrow$ but bias $\uparrow$,
- $\lambda \downarrow \implies$ bias $\downarrow$ but variance $\uparrow$.
- Would like to choose λ to find optimal bias-variance tradeoff:
  - ↪ Unfortunately, optimal λ will generally depend on unknown g!

- Fitted values are $\hat{y} = S_\lambda y$.
- Fitted value $\hat{y}_j^-$ obtained when y_j is dropped from fit is

$$S_{jj}(\lambda)(y_j - \hat{y}_j^-) = \hat{y}_j - \hat{y}_j^- .$$

- Cross-validation sum of squares is

$$\text{CV}(\lambda) = \sum_{j=1}^{n}(y_j - \hat{y}_j^-)^2 = \sum_{j=1}^{n}\left\{\frac{y_j - \hat{y}_j}{1 - S_{jj}(\lambda)}\right\}^2,$$

and generalised cross-validation sum of squares is

$$\text{GCV}(\lambda) = \sum_{j=1}^{n}\left\{\frac{y_j - \hat{y}_j}{1 - \text{trace}(S_\lambda)/n}\right\}^2,$$

where $S_{jj}(\lambda)$ is (j, j) element of S_λ.

Depending on what $\mathcal{F} \ni g(\cdot)$ is (Hilbert space) can write:

$$g(x) = \sum_{k=1}^{\infty} \beta_k \psi_k(x) \quad \text{(in an appropriate sense)},$$

with $\{\psi\}_{k=1}^{\infty}$ known (orthogonal) basis functions for $\mathcal{F}$, e.g.,

- $\mathcal{F} = L^2(-\pi, \pi)$,
- $\{\psi_k\} = \{e^{-ikx}\}_{k \in \mathbb{Z}}$, $\psi_i \perp \psi_j$, $i \neq j$.
- Gives Fourier series expansion, $\beta_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) e^{-ikx} \, dx$.

Idea: if truncate series, then have simple linear regression!

$$Y_i = \sum_{k=1}^{\tau} \beta_k \psi_k(x_i) + \varepsilon_i, \quad \tau < \infty$$

Notice: truncation has implications, e.g., in Fourier case:

- Truncating implies assume $g \in \mathcal{G} \subset L^2$.
- Interpret this as a smoothness assumption on g.
- How to choose τ optimally?

Easy exercise in Fourier analysis:

$$\sum_{k=-\tau}^{\tau} \beta_k \, e^{-ikx} = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(y) D_\tau(x-y) \, dy$$

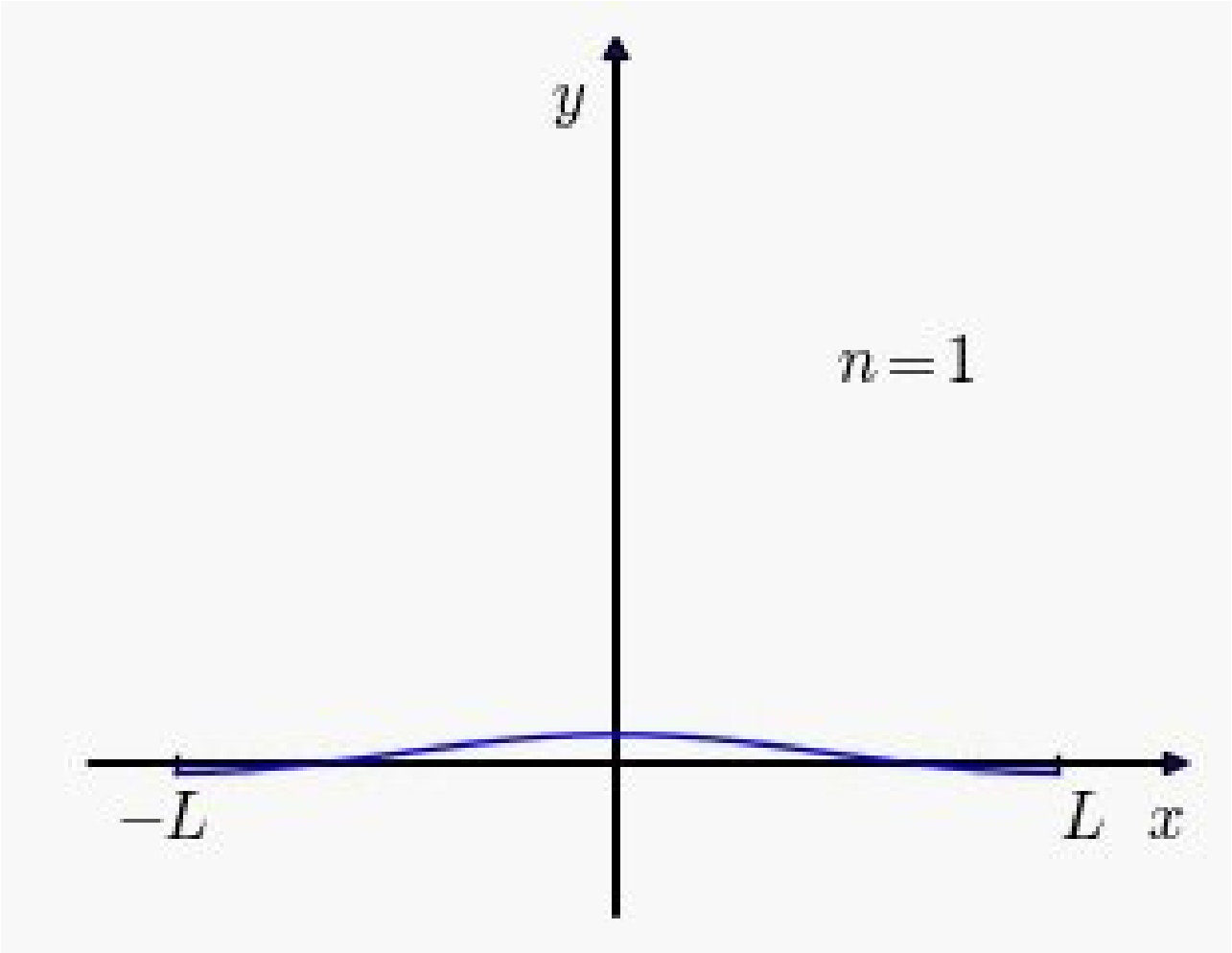
with the *Dirichlet kernel* of order τ, $D_\tau(u) = \sin\{(\tau + 1/2)\, u\}/\sin(u/2)$.
Recall kernel smoother:

$$\hat{g}(x_0) = \sum_{i=1}^{n} \frac{y_i \, K_\lambda(x_i - x_0)}{\sum_{i=1}^{n} K_\lambda(x_i - x_0)} = \frac{1}{c} \int_I y(x) K_\lambda(x - x_0) \, dx,$$

with

$$y(x) = \sum_{i=1}^{n} y_i \delta(x - x_i).$$

- So if K is the Dirichlet kernel, we can do series approximation via kernel smoothing.
- Works for other series expansions with other kernels (e.g., Fourier with convergence factors)


y
$n=1$
$-L$
L
x

So far: how to estimate $g : \mathbb{R} \to \mathbb{R}$ (assumed smooth) in

$$Y_i = g(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \text{ given data } \quad \{(y_i, x_i)\}_{i=1}^n.$$

- Generalise to include multivariate explanatories?
- "Immediate" Generalisation: $g : \mathbb{R}^p \to \mathbb{R}$ (smooth)

$$Y_j = g(x_{j1}, \ldots, x_{jp}) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

- Estimation by (e.g.) multivariate kernel method.
- Two basic drawbacks of this approach ...

↪ Shape of kernel? (definition of *local*)

↪ *Curse of dimensionality*

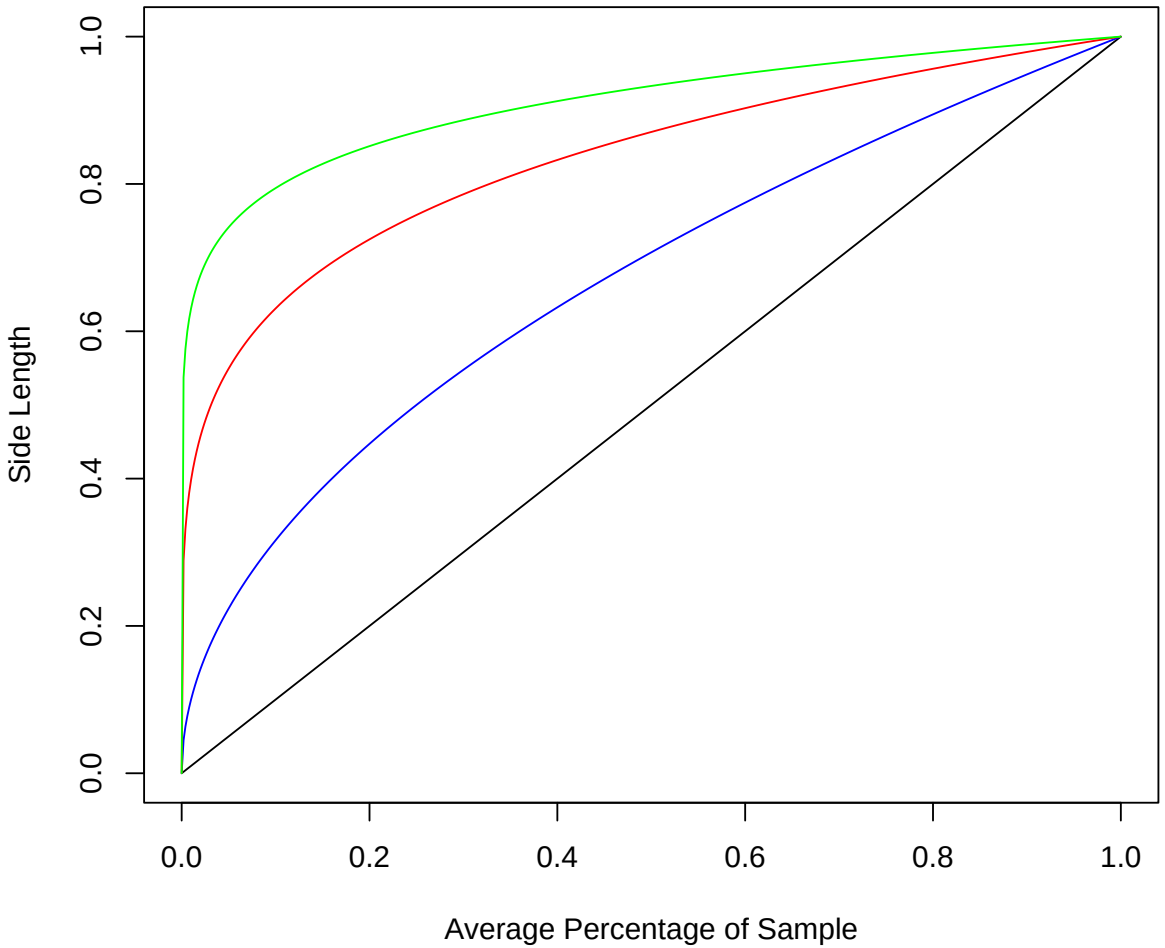
$\rightarrow$ Need some definition of "local" in the space of explanatories

$\hookrightarrow$ Use some metric on $\mathbb{R}^p \ni (x_1, \ldots, x_p)$!

But which one?

- Choice of metric $\iff$ choice of geometry
  - $\hookrightarrow$ e.g., curvature reflects intertwining of dimensions
- Geometry $\implies$ reflects structure in the explanatories
  - potentially different units of measurement / (variable stretching of space)
  - g may be of higher variation in some dimensions / (need finer neighbourhoods there)
  - statistical dependencies present in the explanatories / ("local" should reflect these)

Curse of Dimensionality ($\mathcal{U}[0,1]^p$)



$p = 1$, $p = 2$, $p = 5$, $p = 10$

"neighbourhoods with a fixed number of points become less local as the dimensions increase"

Bellman (1961)

- Notion of local in terms of % of data: fails in high dimensions
  $\hookrightarrow$ There is too much space!
- Hence to allow for reasonably small bandwidths
  $\hookrightarrow$ Density of sampling must increase.
- Need to have ever larger samples as dimension grows.

Attempt to find a link/compromise between:

- our mastery of 1D case (at least we can do that well . . .),
- and higher dimensional explanatories (and associated difficulties).

One approach: Projection-Pursuit Regression

$$Y = \sum_{k=1}^{K} h_k(\vartheta_k^\top \mathbf{x}) + \varepsilon, \quad \|\vartheta_k\| = 1,\ \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- Additively decomposes g into smooth functions $h_k : \mathbb{R} \to \mathbb{R}$.
- Each function depends on a global feature
  $\hookrightarrow$ a linear combination of the explanatories,
- projections directions chosen for best fit
  $\hookrightarrow$ similarities to tomography.
- Each h_k is a ridge function of $\mathbf{x}$: varies only in the direction defined by ϑ_k

How is the model fitted to data?

Assume only one term, $K = 1$ and consider penalized likelihood:

$$\min_{h \in C^2, \|\vartheta\|=1} \left\{ \sum_{i=1}^{n} \{y_i - h_1([\vartheta^\top \mathbf{x}]_i)\}^2 + \int_I \{h_1''(t)\}^2 \, dt \right\}.$$

Two steps:

- *Smooth*: Given a direction ϑ, fitting $g_1(\vartheta^\top \mathbf{x})$ is done via 1D smoothing splines.
- *Pursue*: Given h_1, have a non-linear regression problem w.r.t. ϑ.

Hence, iterate between the two steps

↪ Complication is that h_1 not explicitly known, so need numerical derivatives.

↪ Computationally intensive (impractical in the '80's).

→ Further terms added in forward stepwise manner.

Projection pursuit:

(+) Can uniformly approximate $C^1(\text{compact}[\mathbb{R}^p])$ function arbitrarily well as $K \to \infty$ (very useful for prediction)

(−) Interpretability? What do terms mean within problem?

Need something that can be interpreted variable-by-variable

- Compromise: Additive Model

$$Y_j = \alpha_j + \sum_{k=1}^{p} f_k(x_{jk}) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

  - f_j's univariate smooth functions, $\sum_j f_k(x_{jk}) = 0$.

In our standard setting, have:

$$Y_j \mid \widetilde{x}_j \overset{ind}{\sim} \text{Dist}(\cdot \mid \theta_j) \to \begin{cases} \text{Dist} = \mathcal{N}(\mu_j, \sigma^2), \\ \theta_j = \mu_j = \alpha_j + \sum_{k=1}^{p} f_k(x_{jk}). \end{cases}$$

- How to fit additive model?

↪ Know how to fit each f_k separately quite well

↪ Take advantage of this ...

- Motivation: Fix j and drop it for ease:

$$\mathbb{E}\left[Y - \alpha - \sum_{m \neq k} f_m(x_m)\right] = f_k(x_k)$$

- Suggests the *Backfitting Algorithm*:

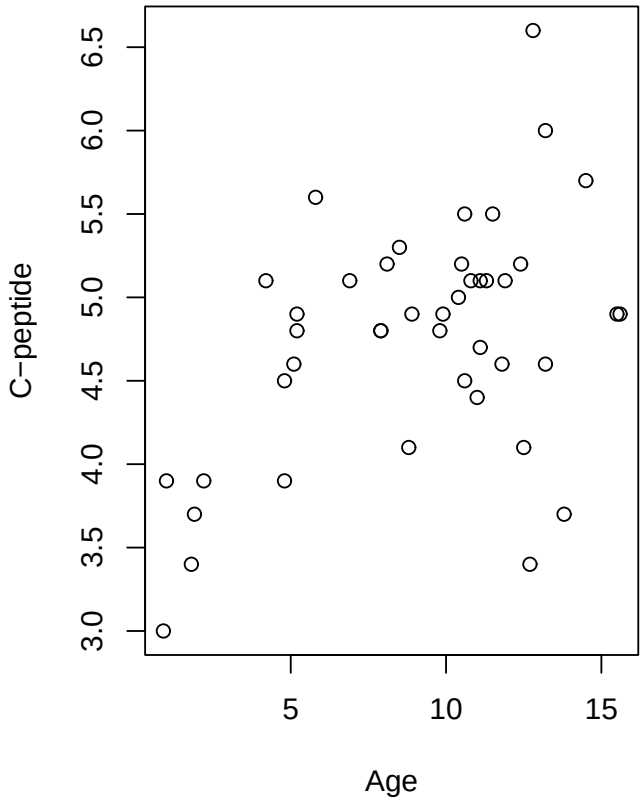
(1) `Initialise:` $\alpha = \text{ave}\{y_j\}$, $f_k = f_k^0$, $k = 1, \ldots, p$.

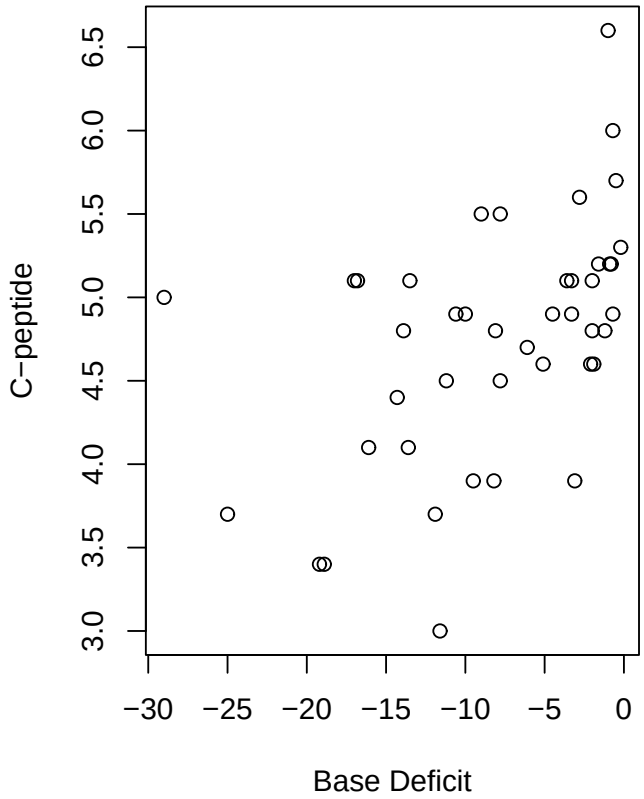
(2) `Cycle:` $f_k = \mathcal{S}_k(y - \alpha - \sum_{m \neq k} \mathbf{f}_m)$ $k = 1, \ldots, p, 1, \ldots, p, \ldots$

(3) `Stop:` when individual functions don't change

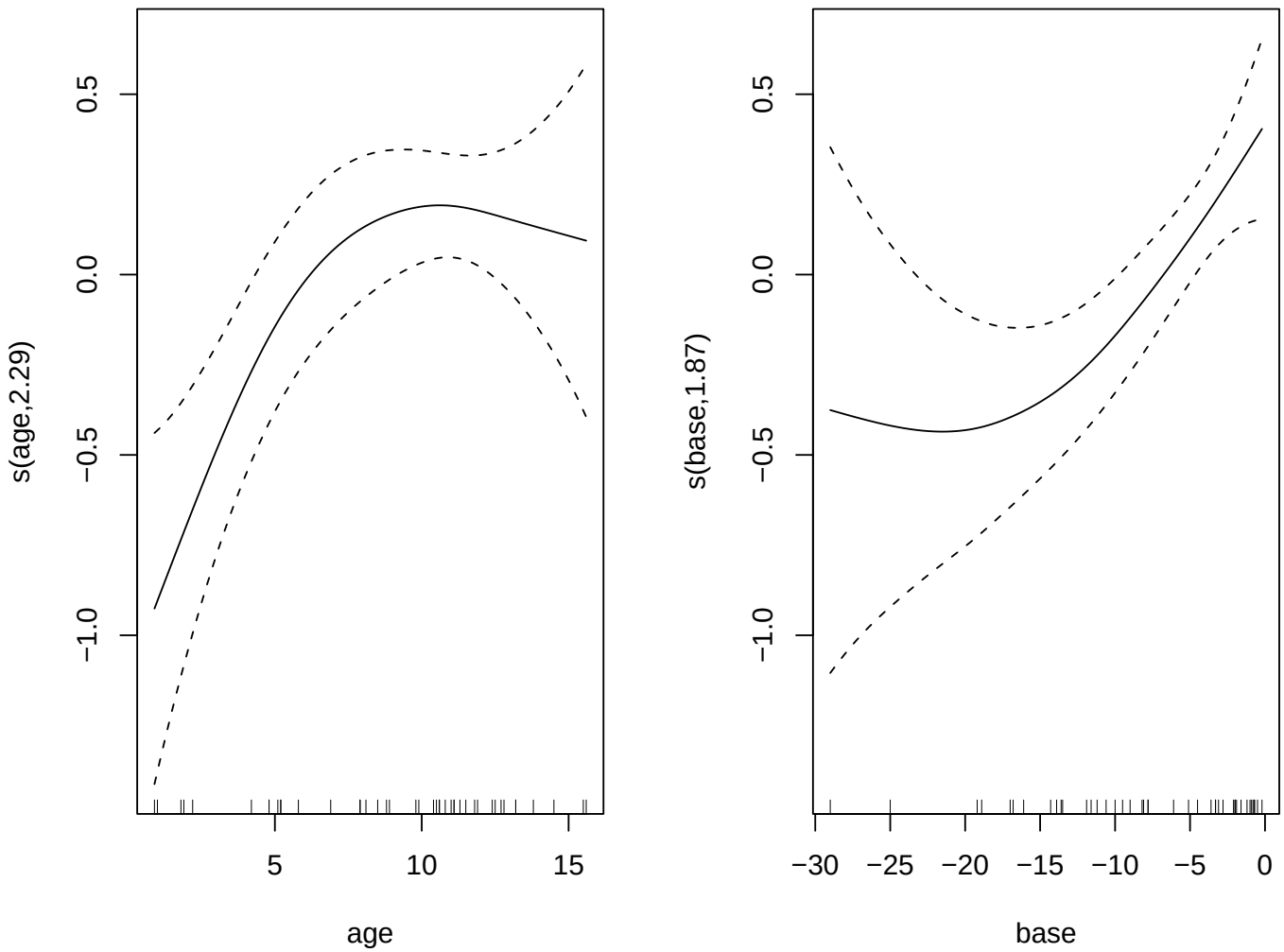
- $\mathcal{S}$ is arbitrary scatterplot smoother

Example: Diabetes Data



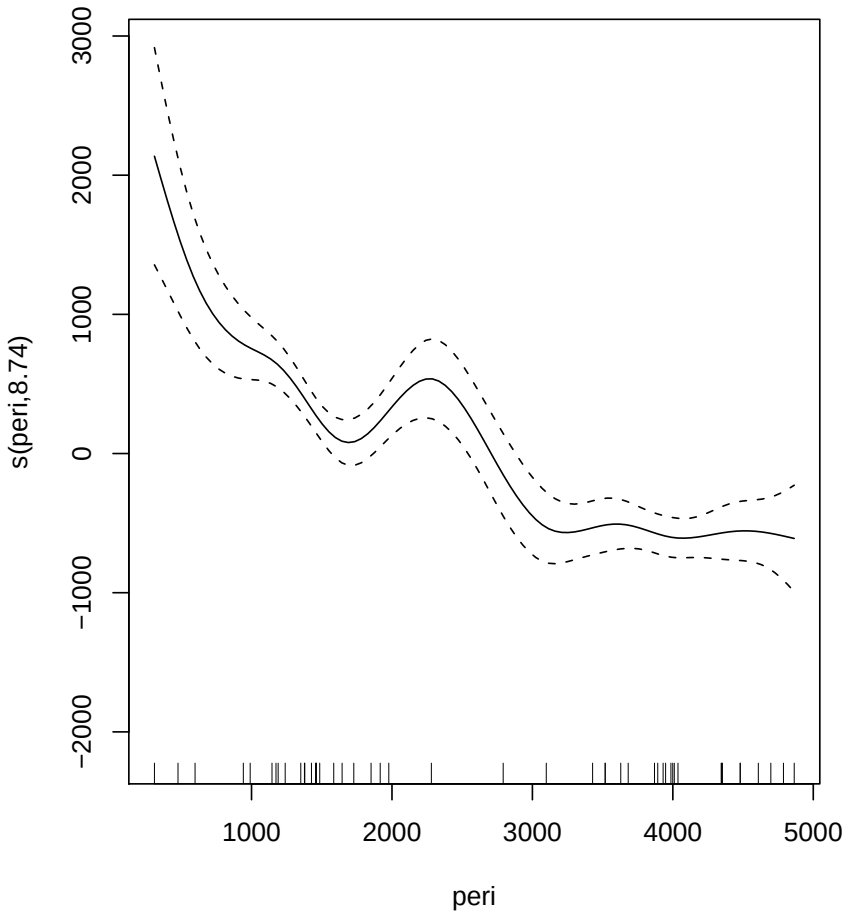


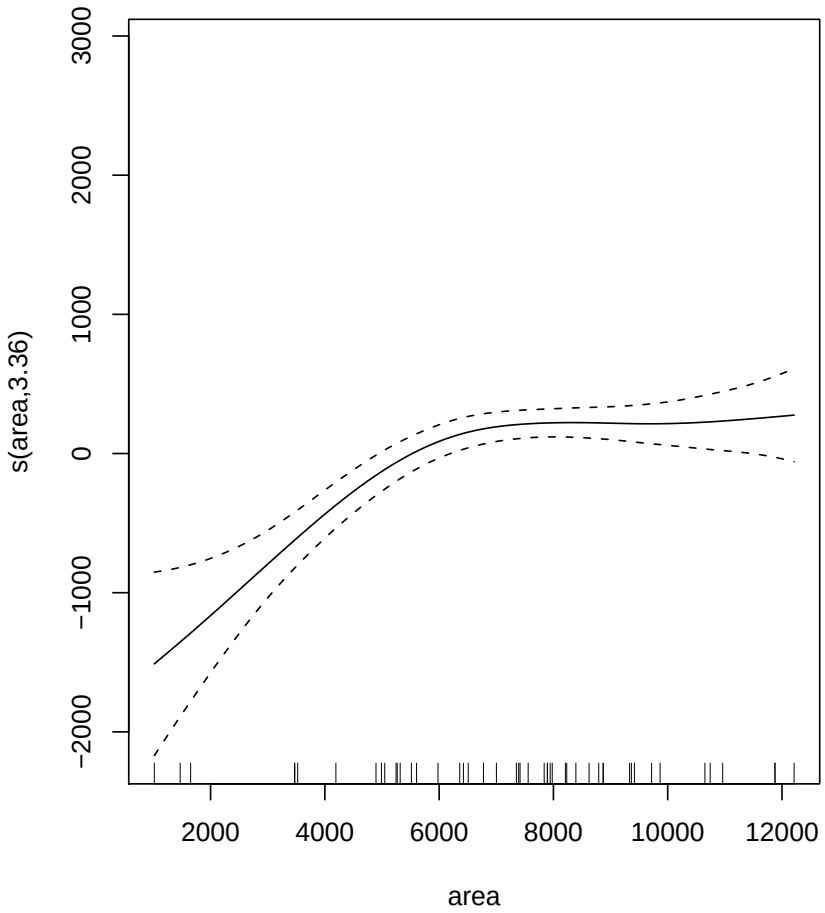
Example: Diabetes Data



Measurements on 48 rock samples from a petroleum reservoir:

```
rock.gam<-gam(perm 1+s(peri)+s(area),family=gaussian)
```





```
Family: gaussian
Link function: identity

Formula:
perm ~ 1 + s(peri) + s(area)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.45      27.18   15.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
          edf Est.rank      F  p-value
s(peri) 8.739        9 18.286 9.49e-11 ***
s(area) 3.357        7  6.364 7.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.815   Deviance explained = 86.3%
```

More on Splines

We want to rigorously show:

1. The penalized least squares problem admints a natural cubic spline as a unique solution
2. That any natural cubic spline on n distinct knots can be expanded in a basis of n elements $\{B_1, ..., B_n\}$
3. That the matrix inversion involved in the expression $(B^\top B + \lambda\Omega)^{-1} B^\top y$ is well-defined

En route, we would also like to

4. Construct at least one example of an explicit basis $\{B_1, ..., B_n\}$.

To analyse spline smoothing we will need to first analyse spline interpolation.

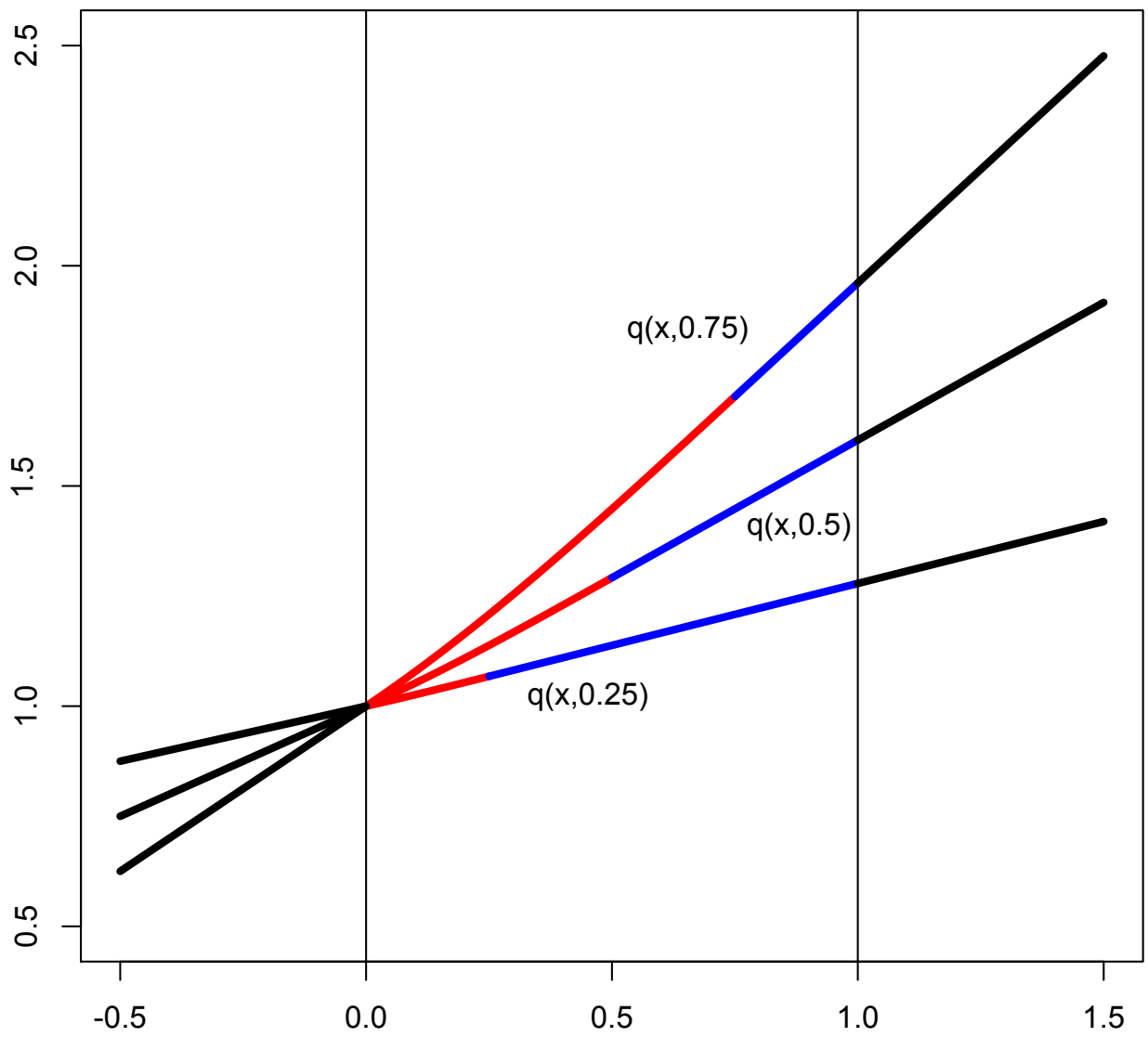
Our analysis will hinge on a very carefully chosen *kernel*:

$$q(x,y) = 1 + xy + k(x,y), \quad (x,y) \in [0,1]^2,$$

where

$$k(x,y) = \begin{cases} x^2 y/2 - x^3/6 & \text{for } x \leq y \\ xy^2/2 - y^3/6 & \text{for } x > y \end{cases}, \quad (x,y) \in [0,1]^2.$$

- We will write $q_y(x)$ or $k_y(x)$ whenever we want to emphasise that the second argument is taken fixed and we view the kernel as a function of the first argument (note both kernels are symmetric).
- In this light, $q_y(x)$ is piecewise polynomial with two pieces:
  1. a cubic piece (for x betweenn 0 and y), and
  2. a linear piece (for x between y and 1).



Recall, $q_y(x)$ is piecewise polynomial with two pieces:

1. a cubic piece (for $0 \leq x \leq y$)
2. a linear piece (for $y \leq x \leq 1$).

Theorem (Positive Definiteness)

Given any $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n \leq 1$ *we have*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \, q(t_i, t_j) \geq 0 \qquad \forall\, \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^\top \in \mathbb{R}^n,$$

in other words $\boldsymbol{Q} = \{q(t_i, t_j)\}_{i,j=1}^n$ *is nonnegative definite. When all the* t_j*'s are distinct,*

$$0 \leq t_1 < t_2 < \ldots < t_n \leq 1,$$

the displayed inequality is strict unless $\boldsymbol{\alpha} = 0$*, and so* $\boldsymbol{Q}$ *is positive definite.*

Proof.

Let $\boldsymbol{K} = \{k(t_i, t_j)\}_{i=1}^n$, $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$, $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$ and note that

$$\boldsymbol{Q} = \{q(t_i, t_j)\}_{i,j=1}^n = \{1 + t_i t_j + k(t_i, t_j)\}_{i,j=1}^n = \mathbf{1}\mathbf{1}^\top + \boldsymbol{t}\boldsymbol{t}^\top + \boldsymbol{K}.$$

Thus, if we can verify that $\boldsymbol{K} \succeq 0$ we will obtain that $\boldsymbol{Q} \succeq 0$, being the sum of three non-negative definite matrices.

Given any pair (t_i, t_j) with $t_i \leq t_j$ (say), observe that

$$\int_0^1 k_{t_i}''(u) k_{t_j}''(u)\, du = \int_0^{t_i} (t_j - u)(t_i - u)\, du = t_i^2 t_j/2 - t_i^3/6 = k(t_i, t_j). \quad (*)$$

Therefore, we may substitute the integral expression for $k(t_i, t_j)$ into $\boldsymbol{\alpha}^\top \boldsymbol{Q} \boldsymbol{\alpha}$ to manifest a square:

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(t_i, t_j) = \sum_{i=1}^n \sum_{j=1}^n \int_0^1 \alpha_i k_{t_i}''(u) \alpha_j k_{t_j}''(u)\, du = \int_0^1 \left(\sum_{i=1}^n \alpha_i k_{t_i}''(u) \right)^2 du.$$

This shows that $\boldsymbol{\alpha}^\top \boldsymbol{Q} \boldsymbol{\alpha} \geq 0$, and so $\boldsymbol{K}$ (and hence $\boldsymbol{Q}$) is always nonnegative.

Now suppose that the $\{t_i\}$ are all distinct. Remark that each function $k_{t_i}''(u)$ is supported on $[0, t_i)$ and is linear thereon. We distinguish two cases:

- $t_1 > 0$. Then all n supports are disjoint non-empty intervals and the $\{k_{t_j}''\}_{j=1}^n$ are linearly independent. Consequently the sum can be zero only if $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$, and $\boldsymbol{K}$ (and hence $\boldsymbol{Q}$) is strictly positive.

- $t_1 = 0$. Then $k''_{t_1} = 0$, so only the $n-1$ functions $\{k''_{t_j}\}_{j=2}^n$ are linearly independent. In this case, first row/column of $\boldsymbol{K}$ will be uniformly zero, and only the bottom right $(n-1) \times (n-1)$ submatrix

$$\boldsymbol{K}_{n-1} = \{k(t_i, t_j)\}_{j=2}^n$$

of $\boldsymbol{K}$ will be positive definite. Thus $\boldsymbol{K}$ is of reduced rank $n-1$. However, the first column of $\mathbf{1}\mathbf{1}^\top$ is now linearly independent of all columns of $\boldsymbol{K}$, and so $\boldsymbol{Q} = \mathbf{1}\mathbf{1}^\top + \boldsymbol{t}\boldsymbol{t}^\top + \boldsymbol{K}$ is of full rank n.

In summary, when $0 \leq t_1 < t_2 < \ldots < t_n \leq 1$, the matrix $\boldsymbol{Q}$ is positive definite.

□

Notice that the calculation $(*)$ was the crucial ingredient. We will use this again when proving that $\boldsymbol{\Omega}$ is nonnegative.

Why go into all this trouble? It turns out that this property will give us both:

- A solution to the spline interpolation problem.
- A basis for natural cubic splines.

Theorem (Spline Interpolation: Uniqueness and Optimality))

Let $0 = t_1 < t_2 < \ldots < t_n = 1$ *be distinct nodes, with* $n \geq 2$, *and* $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ *be associated responses.*

1. *There exists a unique natural cubic spline* $s : [0,1] \to \mathbb{R}$ *with knots at* $\{t_j\}$ *that interpolates* $\{(t_j, y_j)\}_{j=1}^n$, *and can be explicitly constructed as*

$$s(x) = \sum_{j=1}^{n} \theta_i \, q(x, t_j), \qquad \text{with } \boldsymbol{\theta} = \boldsymbol{Q}^{-1}\boldsymbol{y}$$

   where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$, *and* $\boldsymbol{Q} = \{q(t_i, t_j)\}_{i,j=1}^n$ *is bone fide invertible.*

2. *for any* C^2 *function* $f : [0,1] \to \mathbb{R}$ *that also interpolates* $\{(t_j, y_j)\}_{j=1}^n$,

$$\mathcal{C}(f) \equiv \int_0^1 [f''(u)]^2 \, du \geq \int_0^1 [s''(u)]^2 \, du \equiv \mathcal{C}(s). \qquad (\mathsf{I})$$

3. *The inequality in* (I) *is strict unless* $f(u) = s(u)$ *everywhere on* $[0,1]$.

Proof.

Notice that $\boldsymbol{Q}$ is indeed invertible by our previous theorem, so $s(x)$ is well-defined and indeed a natural cubic spline by definition.
To verify that it interpolates $\{(t_j, y_j)\}_{j=1}^n$, write $\boldsymbol{s} = (s(t_1), \dots, s(t_n))^\top$ and note

$$s(t_i) = \sum_{j=1}^n \theta_i q(t_i, t_j), \quad \text{and so} \quad \boldsymbol{s} = \boldsymbol{Q\theta} = \boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{y} = \boldsymbol{y}.$$

This establishes existence of at least one interpolating cubic spline, constructible explicitly via the stated form. To establish that this is the unique interpolating spline, we will:

- prove that (2) and (3) hold for any interpolating spline (not s specifically).
- using this, we will show that there can only be one interpolating spline

thus closing our proof loop.

Let f be an arbitrary C^2 interpolant and $w(x)$ be an interpolating cubic spline, not necessarily equal to s. Define $\delta(x) = f(x) - w(x)$ and remark that $\delta(t_j) = 0$ for all j since w interpolates f at the nodes. Now, expand the square to write

$$\mathcal{C}(f) = \mathcal{C}(w + \delta) = \mathcal{C}(w) + \mathcal{C}(\delta) + \int_a^b w''(u)\delta''(u)du.$$

We claim that the last term vanishes. Using integration by parts

$$\int_a^b w''(u)\delta''(u)\,du = w''\delta'\Big|_0^1 - \int_0^1 w'''(u)\delta'(u)\,du = -\int_0^1 w'''(u)\delta'(u)\,du$$

because $w''(0) = w''(1) = 0$ by the natural boundary constraint. Breaking the integration over the knot partition and using integration by parts a second time,

$$\int_0^1 w'''(u)\delta'(u)\,du = \sum_{j=1}^{n-1} \int_{t_j}^{t_j+1} w'''(u)\delta'(u)\,du =$$

$$= \sum_{j=1}^{n-1} \left(w'''\delta\Big|_{t_j}^{t_j+1} - \int_{t_j}^{t_j+1} w''''(u)\delta(u)\,du \right) = 0$$

because on each partition subinterval w''' is a constant and w'''' vanishes, whereas $\delta(t_j) = 0$ by the interpolation constraint.

This establishes that for any C^2 interpolant f and any interpolating natural cubic spline w, we must have

$$\mathcal{C}(f) = \mathcal{C}(w) + \mathcal{C}(\delta) \geq \mathcal{C}(w).$$

The inequality

$$\mathcal{C}(f) = \mathcal{C}(w) + \mathcal{C}(\delta) \geq \mathcal{C}(w).$$

becomes an equality if and only if $\mathcal{C}(\delta) = 0$. But if $\mathcal{C}(\delta) = 0$, it must be that $\delta'' = 0$ because δ'' is continuous (by w'' and f'' being so). Hence, δ is linear everywhere on $[0, 1]$, and so must be uniformly zero on $[0, 1]$ since $\delta(t_j) = 0$.

In summary, for any interpolating spline w and any C^2 interpolant,

$$\mathcal{C}(f) \geq \mathcal{C}(w), \quad \text{unless } f = w. \tag{C}$$

Let us use this conclusion to establish uniqueness in (1). Let $s_1(x)$ and $s_2(x)$ be two natural cubic splines that interpolate $\{(t_j, y_j)\}_{j=1}^n$. Apply conclusion (C) to s_1 and s_2 twice, each time reversing their roles:

- First, take s_2 as an interpolating spline and s_1 as some C^2 interpolant. We must have $\mathcal{C}(s_1) > \mathcal{C}(s_2)$ unless $s_1 = s_2$.
- Second, take s_1 as an interpolating spline and s_2 as some C^2 interpolant. We must have $\mathcal{C}(s_2) > \mathcal{C}(s_1)$ unless $s_2 = s_1$.

The only way for the two conclusions to hold simultaneously is for $s_1 = s_2$, which proves uniqueness in (1) and completes the proof. □

Corollary

Given distinct nodes $0 = t_1 < t_2 < \ldots < t_n = 1$, *the set* $\mathcal{S}(t_1, ..., t_n)$ *of natural cubic splines with knots* $\{t_j\}_{j=1}^n$ *is a vector space of dimension* n, *and*

$$\varphi_i(x) = q_{t_i}(x) = q(x, t_i), \qquad i = 1, ..., n$$

forms a basis for $\mathcal{S}(t_1, ..., t_n)$.

Proof.

It is immediate that $\mathcal{S}(t_1, ..., t_n)$ is a vector space by the definition of a natural cubic spline. And, for any $\boldsymbol{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$ there is a unique $s_{\boldsymbol{y}} \in \mathcal{S}(t_1, ..., t_n)$ that interpolates $\{(t_j, y_j)\}_{j=1}^n$. This establishes a bijection between $\mathbb{R}^n$ and $\mathcal{S}(t_1, ..., t_n)$, and proves that the dimension of $\mathcal{S}(t_1, ..., t_n)$ is n.

To show that the collection of n functions $\{\varphi_i\}_{i=1}^n$ is linearly independent, we need to show that if $\theta_1\varphi_1(x) + \ldots \theta_n\varphi_n(x) = 0$, then $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top = \mathbf{0}$

Note that $0 = \sum_{j=1}^n \theta_j \varphi_j(x) \equiv \sum_{j=1}^n \theta_j\, q_j(x, t_j)$, is the unique natural cubic spline that interpolates $\{(t_j, 0)\}_{j=1}^n$. Hence, we must have that $\boldsymbol{Q\theta} = \mathbf{0}$, for $\boldsymbol{Q} = \{q(t_i, t_j)\}_{i,j=1}^n$ strictly positive definite, and so $\boldsymbol{\theta} = 0$. □

Corollary

If $\{B_i\}_{i=1}^n$ is a basis for natural cubic splines on n distinct nodes $0 = t_1 < \ldots < t_n = 1$, then the $n \times n$ matrix $\boldsymbol{B} = \{B_i(t_j)\}_{i,j=1}^n$ is invertible and the $n \times n$ matrix $\boldsymbol{\Omega} = \{\int_0^1 B_m''(x)B_k''(x)dx\}_{m,k=1}^n$ is nonnegative definite.

Proof.

The matrix $\boldsymbol{B}$ is invertible if and only if the equation

$$\boldsymbol{B}\boldsymbol{\gamma} = \boldsymbol{y}$$

has a unique solution with respect to $\boldsymbol{\gamma} \in \mathbb{R}^n$ for any $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Notice, however, that as

$$\left\{\sum_{i=1}^n \gamma_i B_i : \boldsymbol{\gamma} \in \mathbb{R}^n\right\} = \mathcal{S}(t_1, ..., t_n)$$

since $\{B_i\}$ is a basis of $\mathcal{S}$. Hence the matrix statement is equivalent to asking whether for any $\boldsymbol{y}$, there exists a unique $s \in \mathcal{S}(t_1, ..., t_n)$ such that

$$s(t_j) = y_j, \qquad j = 1, ..., n.$$

This is guaranteed by the unique interpolation theorem.

To show $\boldsymbol{\Omega} \succeq 0$, note that each B_m can be expanded in the basis $\{q_{t_i}(x)\}_{i=1}^n$ as

$$B_m(x) = \sum_{i=1}^n \theta_{i,m}\, q_{t_i}(x).$$

Therefore,

$$B_m''(x) = \sum_{i=1}^n \theta_{i,m}\, q_{t_i}''(x) = \sum_{i=1}^n \theta_{i,m}\, k_{t_i}''(x), \quad m = 1, \ldots, n.$$

Consequently, we can make use of our earlier calculation $(*)$ to get

$$\int_0^1 B_m''(x) B_k''(x)\, dx = \sum_{i=1}^n \sum_{j=1}^n \theta_{i,m}\theta_{j,k} \int_0^1 k_{t_i}''(x) k_{t_j}''(x)\, dx =$$

$$\overset{(*)}{=} \sum_{i=1}^n \theta_{i,m} \sum_{j=1}^n k(t_i, t_j)\theta_{j,k}.$$

Equivalently, $\boldsymbol{\Omega} = \boldsymbol{\Theta}^\top \boldsymbol{K} \boldsymbol{\Theta}$, for $\boldsymbol{\Theta} = \{\theta_{i,m}\}_{i,m=1}^n$ so $\boldsymbol{\Omega} \succeq 0$. □.

Theorem (Splines Minimise Penalised Least Squares)

Given covariates $0 = x_1 < \ldots < x_n = 1$ and responses $\{y_i\}_{i=1}^n$, the functional

$$\mathcal{L}(f) = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \int_0^1 (f''(u))^2 \, du$$

is uniquely minimised at a natural cubic spline $\hat{f}(x)$ with knots $\{x_j\}_{j=1}^n$ expressed as

$$\hat{f}(x) = \sum_{j=1}^{n} \hat{\gamma}_j B_j(x),$$

with

$$(\hat{\gamma}_1, \ldots, \hat{\gamma}_n)^\top = \hat{\gamma} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top \boldsymbol{y},$$

where

- $\{B_j(x)\}_{j=1}^n$ is any basis for natural cubic splines with knots $\{x_j\}_{j=1}^n$
- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$
- $\boldsymbol{B} = \{B_j(x_i)\}_{i,j=1}^n$ is invertible.
- $\boldsymbol{\Omega} = \left\{ \int_0^1 B_i''(t) B_j''(t) \, dt \right\}_{i,j=1}^n$ is non-negative definite.

Proof.

Let $f \in C^2$ be a candidate minimiser, and let $s(x)$ be the unique element of $\mathcal{S}(t_1, ..., t_n)$ that interpolates $\{(t_j, f(t_j))\}_{j=1}^n$. Then,

$$\begin{aligned}
\mathcal{L}(f) &= \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \int_0^1 (f''(u))^2 \, du \\
&= \sum_{i=1}^{n} \big(y_i - s(x_i)\big)^2 + \lambda \int_0^1 (f''(u))^2 \, du \\
&\geq \sum_{i=1}^{n} \big(y_i - s(x_i)\big)^2 + \lambda \int_0^1 (s''(u))^2 \, du \;=\; \mathcal{L}(s).
\end{aligned}$$

with equality only if f is itself a spline. Therefore, minimisation of $\mathcal{L}$ over all of C^2, reduces to minimisation of $\mathcal{L}$ over the vector space $\mathcal{S}(t_1, ..., t_n)$. Since $\{B_1, ..., B_n\}$ is a basis for $\mathcal{S}(t_1, ..., t_n)$, our problem is equivalent to minimising

$$\mathcal{G}(\gamma) = \sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{n} \gamma_i B_i(x_j)\right)^2 + \lambda \int_0^1 \left(\sum_{i=1}^{n} \gamma_i B_i''(u)\right)^2 du$$

over $\gamma = (\gamma_1, ..., \gamma_n)^\top \in \mathbb{R}^n$.

In matrix notation, we want to minimize w.r.t. γ the expression

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma})^\top (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\gamma}.$$

This is a ridge regression problem, and will admit the unique solution

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{B}^\top \boldsymbol{y}$$

provided the matrix $\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega}$ is indeed invertible. This follows from the fact that $\boldsymbol{B}$ is invertible (hence $B^\top B$ is strictly positive definite) and $\boldsymbol{\Omega}$ is nonnegative definite, as per our last corollary. □

Nonparametric Regression and Efficiency Considerations

To get a transparent analysis, we let $(X_i, Y_i)_{i=1}^n$ be iid with

$$Y_i = m(X_i) + \varepsilon_i, \qquad X_i \sim \mathrm{Unif}(0,1),$$

and assume ε_i is independent of X_i, with

$$\mathbb{E}[\varepsilon_i] = 0, \qquad \mathbb{E}[\varepsilon_i^2] = \sigma^2.$$

Note that we took random design points, instead of fixed ones, contrary to previous practice. This is to simplify the analysis.

Assume $m \in L^2[0,1]$, and for simplicity, m satisfies periodic boundary constraints.

The natural basis in which to analyse this problem is the Fourier basis:

$$\phi_1(x) = 1, \quad \phi_{2k}(x) = \sqrt{2}\cos(2\pi kx), \quad \phi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx).$$

Then $\int_0^1 \phi_k(x)\phi_\ell(x)\,dx = \mathbf{1}\{k = \ell\}$, and $\sup_x |\phi_k(x)| \le \sqrt{2}$.

Define the Fourier coefficients and truncated series expansion:

$$\theta_k := \int_0^1 m(x)\phi_k(x)\,dx = \langle m, \phi_k \rangle_{L^2}, \qquad m_N(x) := \sum_{k=1}^{N} \theta_k \phi_k(x).$$

Note m_N is the orthogonal projection of m onto $\mathcal{V}_N := \mathrm{span}\{\phi_1, \ldots, \phi_N\}$, hence

$$\|m - m_N\|_{L^2}^2 = \sum_{k>N} \theta_k^2 \qquad \text{(Parseval)}.$$

Define the empirical Fourier coefficients and empirical truncated series estimator:

$$\hat{\theta}_k := \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_k(X_i) = \langle \boldsymbol{Y}, \phi_k \rangle_{\ell^2}, \qquad \hat{m}_N(x) := \sum_{k=1}^{N} \hat{\theta}_k \phi_k(x).$$

For intuition, imagine if $X_i \approx i/n$ were the nodes of a regular grid.

The series estimator is not the OLS estimator w/ design matrix $\Phi_{ik} = \phi_k(X_i)$

Let Φ be the $n \times N$ design matrix $\Phi_{ik} = \phi_k(X_i)$, and let $Y = (Y_1, \ldots, Y_n)^\top$. The OLS estimator for such a model would be

$$\hat{\theta}^{OLS} = \arg\min_{\theta \in \mathbb{R}^N} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{N} \theta_k \phi_k(X_i) \right)^2 = \left(\frac{1}{n} \Phi^\top \Phi \right)^{-1} \left(\frac{1}{n} \Phi^\top Y \right).$$

Writing the Gram matrix $G_n := \frac{1}{n} \Phi^\top \Phi$, we have

$$\hat{\theta}^{OLS} = G_n^{-1} \hat{\theta}.$$

Hence the series estimator coincides with OLS *iff*

$$G_n = I_N \quad \text{(empirical orthonormality)}.$$

- Exact equality on a regular grid (classical Fourier identity). If $X_i = i/n$ (regular grid) and N is below the Nyquist limit, then discrete Fourier orthogonality implies

$$\frac{1}{n}\Phi^\top \Phi = I_N,$$

so $\hat{\theta}^{OLS} = \hat{\theta}$ *exactly*.

- In the random-design setting, the Gram matrix is unbiased. If $X_i \sim \mathrm{Unif}(0,1)$, then

$$\mathbb{E}[G_n] = I_N,$$

since $\mathbb{E}[\phi_k(X)\phi_\ell(X)] = \int_0^1 \phi_k(x)\phi_\ell(x)\,dx = \mathbf{1}\{k=\ell\}$.

- Asymptotic orthogonality and preservation of rates. As $n \to \infty$ (with N fixed, and also for N growing slowly), G_n becomes close to I_N with high probability, hence $G_n^{-1} \approx I_N$. In fact, one can show that G_n concentrates sharply enough around I_N that the rates are preserved, so the performance analysis of the OLS estimator follows from that of the series estimator.

Lemma (Exact (as opposed to *mean*) squared error decomposition)

For any function $g \in \mathcal{V}_N$ and any $m \in L^2[0,1]$,

$$\|g - m\|_{L^2}^2 = \|g - m_N\|_{L^2}^2 + \|m_N - m\|_{L^2}^2,$$

because $g - m_N \in \mathcal{V}_N$ is orthogonal to $m_N - m \in \mathcal{V}_N^{\perp}$.

Apply with $g = \hat{m}_N$:

$$\|\hat{m}_N - m\|_{L^2}^2 = \|\hat{m}_N - m_N\|_{L^2}^2 + \|m_N - m\|_{L^2}^2.$$

By orthonormality,

$$\|\hat{m}_N - m_N\|_{L^2}^2 = \sum_{k=1}^{N} (\hat{\theta}_k - \theta_k)^2, \qquad \|m_N - m\|_{L^2}^2 = \sum_{k>N} \theta_k^2.$$

Lemma

For each fixed $k \geq 1$,

$$\mathbb{E}[\hat{\theta}_k] = \theta_k.$$

Proof.

$$\mathbb{E}[\hat{\theta}_k] = \mathbb{E}[\, Y \phi_k(X)] = \mathbb{E}\big[(m(X) + \varepsilon)\phi_k(X)\big] = \mathbb{E}[m(X)\phi_k(X)] + \mathbb{E}[\varepsilon \phi_k(X)].$$

By independence $\mathbb{E}[\varepsilon \phi_k(X)] = \mathbb{E}[\varepsilon]\mathbb{E}[\phi_k(X)] = 0$. Since $X \sim \text{Unif}(0,1)$,

$$\mathbb{E}[m(X)\phi_k(X)] = \int_0^1 m(x)\phi_k(x)\, dx = \theta_k.$$

□

So the only source of bias is the truncation.

Uniform variance control

$$\mathrm{var}(\hat{\theta}_k) = \frac{1}{n}\mathrm{var}(Y\phi_k(X)) \leq \frac{1}{n}\mathbb{E}[Y^2\phi_k(X)^2] \leq \frac{2}{n}\left(\|m\|_{L^2}^2 + \sigma^2\right).$$

Proof.

Use the facts that $\mathrm{var}(Z) \leq \mathbb{E}[Z^2]$ and $|\phi_k(X)| \leq \sqrt{2}$ to get

$$\mathrm{var}(\hat{\theta}_k) = \frac{1}{n}\mathrm{var}(Y\phi_k(X)) \leq \frac{1}{n}\mathbb{E}[Y^2\phi_k(X)^2] \leq 2\,\mathbb{E}[Y^2].$$

By independence and our moment conditions,

$$\mathbb{E}[Y^2] = \mathbb{E}[m(X)^2] + \sigma^2 + 2\underbrace{\mathbb{E}[\varepsilon m(X)]}_{=0} = \int_0^1 m(x)^2\,dx + \sigma^2.$$

□

Proposition

Under the model assumptions,

$$\mathbb{E}\|\hat{m}_N - m\|_{L^2}^2 = \sum_{k=1}^{N} \mathrm{Var}(\hat{\theta}_k) + \sum_{k>N} \theta_k^2 \leq \frac{2N}{n}\left(\|m\|_{L^2}^2 + \sigma^2\right) + \sum_{k>N} \theta_k^2.$$

Proof.

From the exact decomposition,

$$\mathbb{E}\|\hat{m}_N - m\|_{L^2}^2 = \sum_{k=1}^{N} \mathbb{E}[(\hat{\theta}_k - \theta_k)^2] + \sum_{k>N} \theta_k^2.$$

By unbiasedness, $\mathbb{E}[(\hat{\theta}_k - \theta_k)^2] = \mathrm{Var}(\hat{\theta}_k)$, and apply the variance bound. □

Let's try to understand the bias term (due to truncation) better.

"Nonparametric modelling" amounts to controlling the tail $N \mapsto \sum_{k>N} \theta_k^2$.

For $s > 0$, define the periodic Sobolev ellipsoid

$$\mathcal{W}(s, r) := \left\{ m \in L^2[0,1] : \sum_{j=1}^{\infty} (2\pi j)^{2s} \left(a_j^2 + b_j^2\right) \le r^2 \right\},$$

where a_j, b_j are the cosine/sine Fourier coefficients of m.

Equivalently, writing $m(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x)$, the ellipsoid condition asks

$$\sum_{k=1}^{\infty} \lambda_k^s \theta_k^2 \le r^2$$

where

$$\lambda_1 := 1, \qquad \lambda_{2j} := (2\pi j)^2, \qquad \lambda_{2j+1} := (2\pi j)^2, \qquad j \ge 1.$$

In particular, $\lambda_k \ge 1$ for all k, and $\lambda_k \asymp k^2$.

Assume $s \in \mathbb{N}$ and m is 1-periodic with real Fourier expansion

$$m(x) = a_0 + \sum_{j=1}^{\infty} \Big(a_j \sqrt{2} \cos(2\pi j x) + b_j \sqrt{2} \sin(2\pi j x) \Big).$$

For each N, differentiate the *truncated* Fourier series m_N term-by-term and define

$$m_N^{(s)}(x) = \sum_{j=1}^{N} (2\pi j)^s \Big(a_{j,s} \sqrt{2} \cos(2\pi j x) + b_{j,s} \sqrt{2} \sin(2\pi j x) \Big),$$

where $a_{j,s}, b_{j,s} \in \{\pm a_j, \pm b_j\}$ (signs depend on s). Note that

$$\int_0^1 \big(m_N^{(s)}(x)\big)^2 \, dx = \sum_{j=1}^{N} (2\pi j)^{2s} \big(a_j^2 + b_j^2\big).$$

So if $\sum_{j=1}^{\infty} (2\pi j)^{2s} (a_j^2 + b_j^2) < \infty$, then $(m_N^{(s)})_{N \geq 1}$ is Cauchy in L^2, and thus converges in L^2 to some $m^{(s)} \in L^2[0,1]$, that we call the s-th weak derivative.

Hence the Sobolev ellipsoid condition says we have $m^{(s)} \in L^2$ and $\|m^{(s)}\|_{L^2}^2 \leq r^2$.

Our Sobolev ellipsoid condition, now allows us to quantify the bias term:

Lemma (Tail control)

If $m \in \mathcal{W}(s, r)$, then there exists $C_s > 0$ such that for all $N \geq 1$,

$$\sum_{k>N} \theta_k^2 \leq C_s \, r^2 \, N^{-2s}.$$

Proof.

Let λ_k be as before and recall that $\lambda_k \asymp k^2$. For $k > N$, $\lambda_k^s \geq \lambda_{N+1}^s$, so

$$\sum_{k>N} \theta_k^2 = \sum_{k>N} \lambda_k^{-s} \, (\lambda_k^s \theta_k^2) \leq \lambda_{N+1}^{-s} \sum_{k>N} \lambda_k^s \theta_k^2 \leq \lambda_{N+1}^{-s} \, r^2.$$

Since $\lambda_{N+1}^{-s} \asymp N^{-2s}$, the claim follows. □

Corollary: choose tuning parameter N judiciously to get the rate

Combine the finite-sample bound with the tail lemma: for $m \in \mathcal{W}(s, r)$,

$$\begin{aligned} \mathbb{E}\|\hat{m}_N - m\|_{L^2}^2 &\leq \frac{2N}{n}\left(\|m\|_{L^2}^2 + \sigma^2\right) + C_s r^2 N^{-2s} \\ &\leq \frac{2N}{n}(r^2 + \sigma^2) + C_s r^2 N^{-2s} \end{aligned}$$

where the second inequality is because when $m \in \mathcal{W}(s, r)$,

$$\|m\|_{L^2}^2 = \sum_{k \geq 1} \theta_k^2 \leq \sum_{k \geq 1} \lambda_k^s \theta_k^2 \leq r^2 \qquad \text{(because } \lambda_k \geq 1\text{)}$$

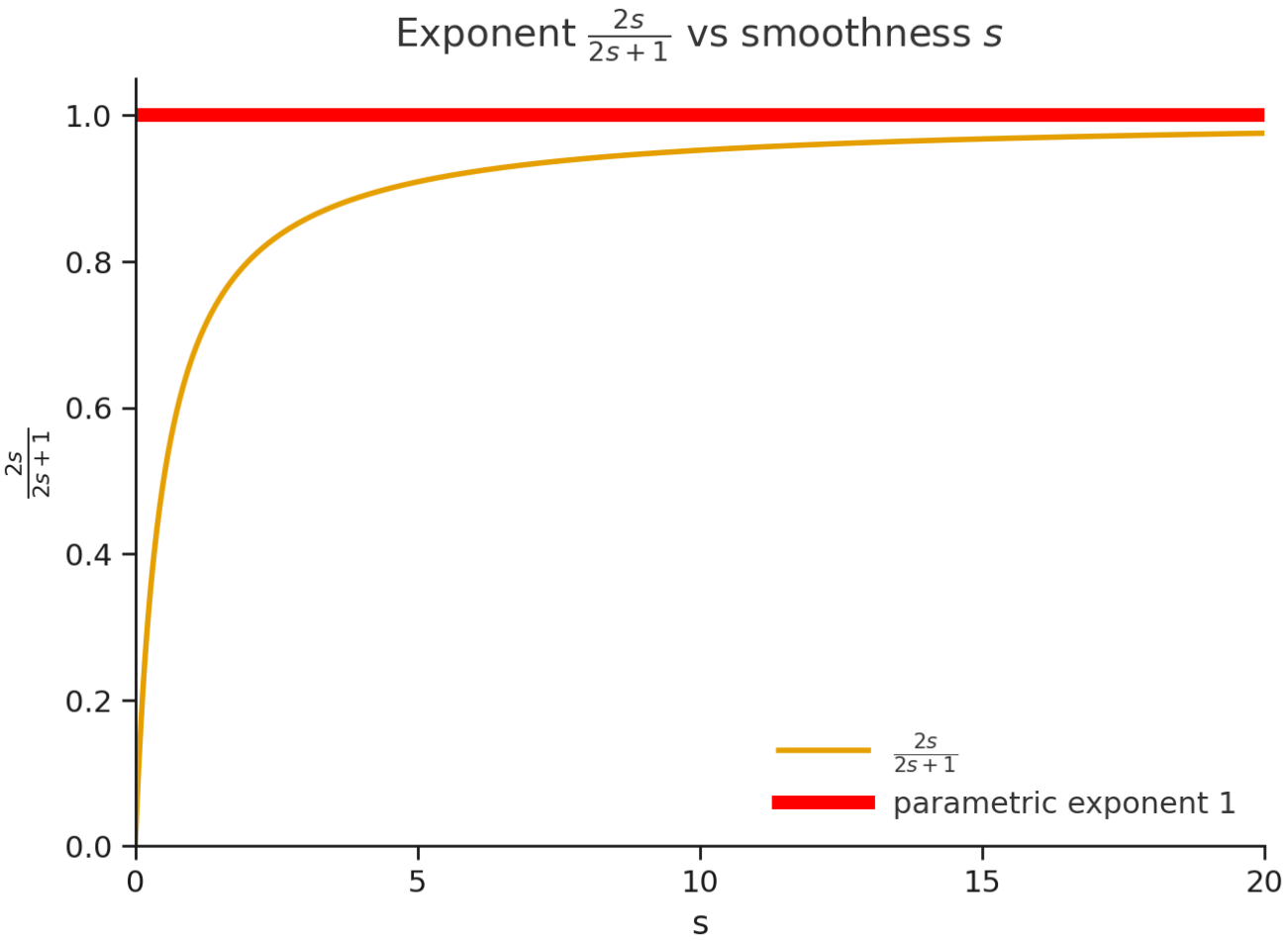
To optimally balance $\frac{N}{n}$ and N^{-2s} we must take $N \asymp n^{\frac{1}{2s+1}}$, giving

$$\sup_{m \in \mathcal{W}(s,r)} \mathbb{E}\|\hat{m}_N - m\|_{L^2}^2 \lesssim n^{-\frac{2s}{2s+1}}.$$

This is a non-asymptotic and uniform upper bound.

Remarkably, we can get get a matching (in order) lower bound over estimators – no measurable map from the data to L^2 can achieve a faster rate.

Nonparametric vs parametric rate



$$n^{-\frac{2s}{2s+1}} \xrightarrow{s\to\infty} n^{-1}$$

To illustrate the effect of dimension cleanly, consider the torus

$$\mathbb{T}^d = \text{“}[0,1]^d \text{ with periodic boundary conditions”}, \qquad X_i \sim \text{Unif}(\mathbb{T}^d).$$

The Fourier basis is $\left\{e^{2\pi i k \cdot x}\right\}_{k \in \mathbb{Z}^d}$, and we write

$$m(x) = \sum_{k \in \mathbb{Z}^d} \theta_k \, e^{2\pi i k \cdot x}.$$

The Sobolev ellipsoid (Fourier definition) now becomes For $s > 0$ and $r > 0$, define

$$\mathcal{W}_d(s, r) := \left\{ m \in L^2(\mathbb{T}^d) : \sum_{k \in \mathbb{Z}^d} (1 + \|k\|^2)^s \, |\theta_k|^2 \leq r^2 \right\}.$$

For integer s, this is (up to constants) equivalent to $\sum_{|\alpha| \leq s} \|\partial^\alpha m\|^2_{L^2(\mathbb{T}^d)} \leq C r^2$.

Keep only frequencies $\|k\| \le M$ (a spectral cutoff).

- Bias (dimension-free exponent). The truncation bias satisfies

$$\|m - m_M\|_{L^2}^2 = \sum_{\|k\|>M} |\theta_k|^2 \le (1+M^2)^{-s} \sum_k (1+\|k\|^2)^s |\theta_k|^2$$
$$\le (1+M^2)^{-s} r^2 \asymp r^2 M^{-2s}.$$

  So the bias exponent is $2s$, independent of d.

- Variance (dimension enters through model size). The number of retained coefficients is

$$\#\{k \in \mathbb{Z}^d : \|k\| \le M\} \asymp M^d.$$

  Estimating each coefficient from n samples costs variance $\asymp 1/n$, hence total variance

$$\text{variance} \asymp \frac{M^d}{n}.$$

Optimal tradeoff achieved at $M \asymp n^{1/(2s+d)}$ and one eventually arrives at

$$\inf_{\tilde{m}} \sup_{m} \mathbb{E}\| \tilde{m} - m \|_{L^2(\mathbb{T}^d)}^2 \asymp n^{-\frac{2s}{2s+d}} \implies \text{curse of dimensionality!}$$

In the same canonical setting as before, we observe iid data

$$(X_i, Y_i)_{i=1}^n, \qquad Y_i = m(X_i) + \varepsilon_i, \qquad X_i \sim \text{Unif}(\mathbb{T}^d),$$

with ε_i independent of X_i, $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.

Additive regression model

Recall that an additive model assumes that the regression function decomposes as

$$m(x) = \sum_{j=1}^{d} m_j(x_j), \qquad x = (x_1, \ldots, x_d) \in \mathbb{T}^d,$$

with the centering convention $\int_0^1 m_j(t)\, dt = 0$ (identifiability).

Idea: instead of estimating one d-dimensional function, we estimate d one-dimensional functions.

Assume each component m_j is 1-periodic and belongs to the 1D Sobolev ellipsoid

$$\mathcal{W}_1(s,r) := \left\{ g \in L^2(\mathbb{T}) : \sum_{\ell \in \mathbb{Z}} (1+\ell^2)^s \, |\vartheta_\ell(g)|^2 \leq r^2 \right\},$$

where $\vartheta_\ell(g)$ denotes the Fourier coefficient of g.

This leads to the additive Sobolev smoothness class (model):

$$\mathcal{A}_d(s,r) := \left\{ m(x) = \sum_{j=1}^{d} m_j(x_j) : \; m_j \in \mathcal{W}_1(s,r), \; \int_0^1 m_j = 0 \right\}.$$

Key points:

- the smoothness index s remains one-dimensional; dimension will only enter through the number of components d.
- this is not just a smootheness based model – not every Sobolev ball function admits this representation – it is a stronger model specificaiton.

For each j, approximate m_j by a 1D Fourier cutoff at frequency M:

$$(m_j)_M(t) = \sum_{|\ell| \le M} \vartheta_\ell(m_j) e^{2\pi i \ell t}.$$

Then

$$m_M(x) := \sum_{j=1}^{d} (m_j)_M(x_j)$$

uses only $d(2M+1) \asymp dM$ coefficients (not M^d!).

- Bias. For each j, $\|m_j - (m_j)_M\|^2_{L^2(\mathbb{T})} \lesssim r^2 M^{-2s}$. By orthogonality/centering cross terms vanish (integrals separate by Fubini),

$$\|m - m_M\|^2_{L^2(\mathbb{T}^d)} \asymp \sum_{j=1}^{d} \|m_j - (m_j)_M\|^2_{L^2(\mathbb{T})} \lesssim d\, r^2 M^{-2s}.$$

- Variance. Estimating $\asymp dM$ coefficients from n samples costs

$$\text{variance} \asymp \frac{dM}{n}.$$

Now we tune the truncation parameter to balance bias/variance:

$$\frac{dM}{n} \asymp d\, M^{-2s} \qquad \Longrightarrow \qquad M \asymp n^{\frac{1}{2s+1}}.$$

Plugging back gives the canonical additive rate

$$\sup_{m \in \mathcal{A}_d(s,r)} \mathbb{E}\| \hat{m} - m \|^2_{L^2(\mathbb{T}^d)} \lesssim d\, n^{-\frac{2s}{2s+1}}.$$

Interpretation

- In the full d-dimensional Sobolev class: $n^{-2s/(2s+d)} \Rightarrow$ exponent deteriorates with d.
- In the additive class: $d\, n^{-2s/(2s+1)} \Rightarrow$ the exponent is 1D, and d appears only as a multiplicative factor.

Additivity relaxes the curse of dimensionality by reducing the effective dimension of the nonparametric part from d to 1.

Takehome messages:

- Tradeoff between flexibility and efficiency
- Parametric model enforces rigid form of parsimony
  (precise formula, few parameters).
- Nonparametric model enforces soft parsimony
  (smoothness, via tuning parameter)
- If model can be confidently assumed, parametric inference is preferable.
- Otherwise, nonparametric methods more flexible and requiring few assumptions.
- Particularly in higher dimensions parametric models more interpretable and efficient.
- But nonparametric curse of dimensionality can be (partially) mitigated by clever approximations (separable/additive models, ridge models,...).
- A very important class of models are semiparametric models. These have some parametric and some nonparametric components.
  - In important cases, can attain parametric efficiency for parametric component.