Regression Methods

Myrto Limnios

Autumn 2025 - Week 1

General Information

- Lectures on Tuesdays, 10h15-12h, CM1221
- Exercises on Tuesdays, 8h15-10h with Marco Piccininni[1]

- Slides and exercises sheets available each Monday evening
- Solutions available on the following Sunday
- Some exercises sessions will be practical and require coding, solutions will be given in `R`
- If you have any question regarding exercises, please come to the dedicated sessions

- No midterm, written final exam only (3h)
- No external material allowed
- We will not respond to emails related to the final exam after December 19th

- **November 18th: Recap session.** No exercise session, problem solving during the lecture session

[1] myrto.limnios@epfl.ch, marco.piccininni@epfl.ch

The Linear Model - Introduction

What is a Regression Model?

Statistical model for:

$$\boxed{y \text{ (random output)} \overset{\text{whose law is influenced by}}{\longleftarrow} x \text{ (non-random input)}}$$

Aim: understand the effect of x on the distribution of random variable Y

General formulation[2]:

$$y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

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

- Inference
- Prediction
- Data compression (parsimonious representations)

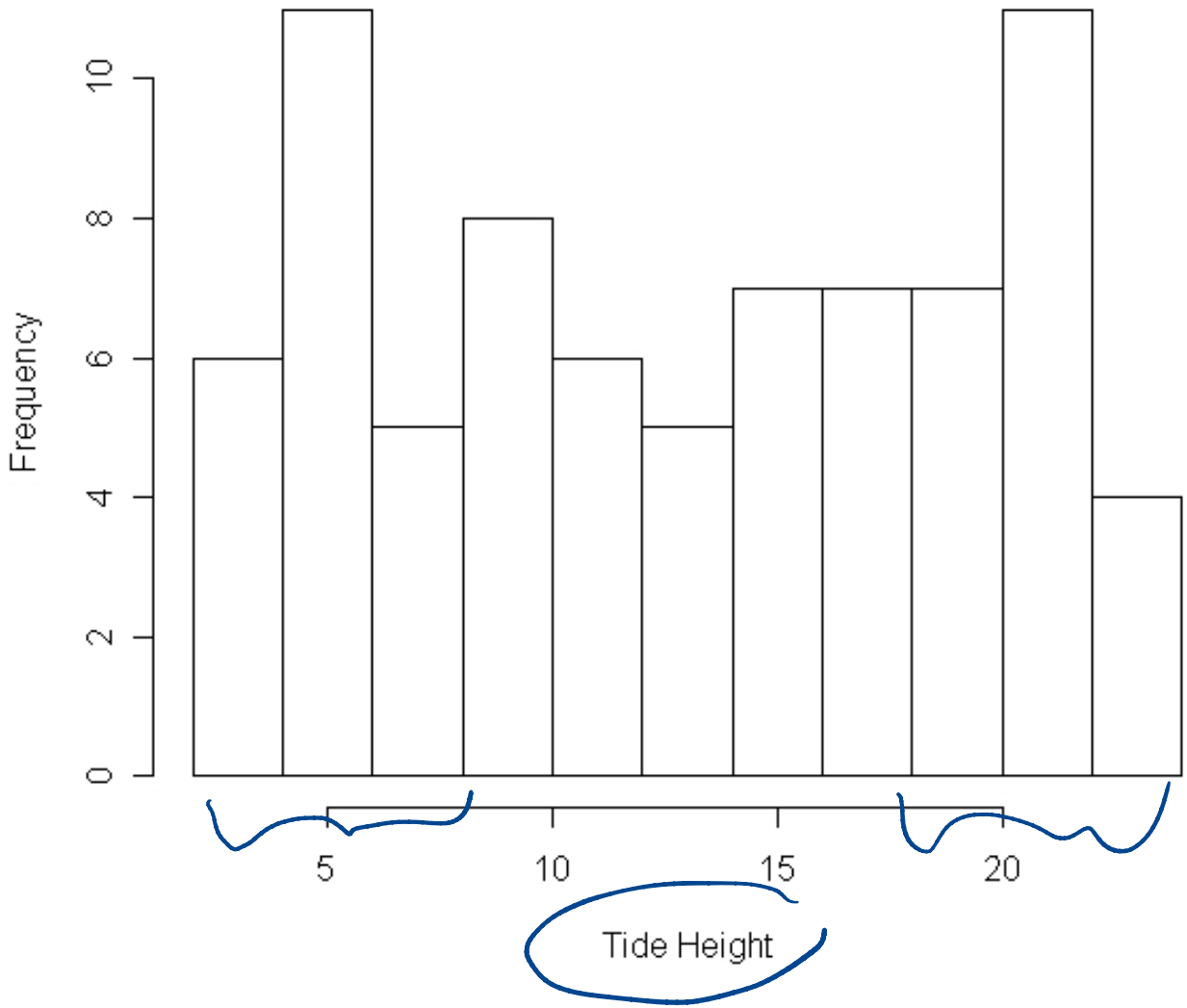
[2] Sometimes we write $y_i | x_i \overset{\text{independent}}{\sim} \text{Distribution}\{g(x_i) = \theta_i\}$ to highlight that the distribution of y depends on x, but without meaning that (x, y) are jointly random; such an assumption is unnecessary (e.g., in a designed experiment we choose values for x).

Absence or presence of covariates

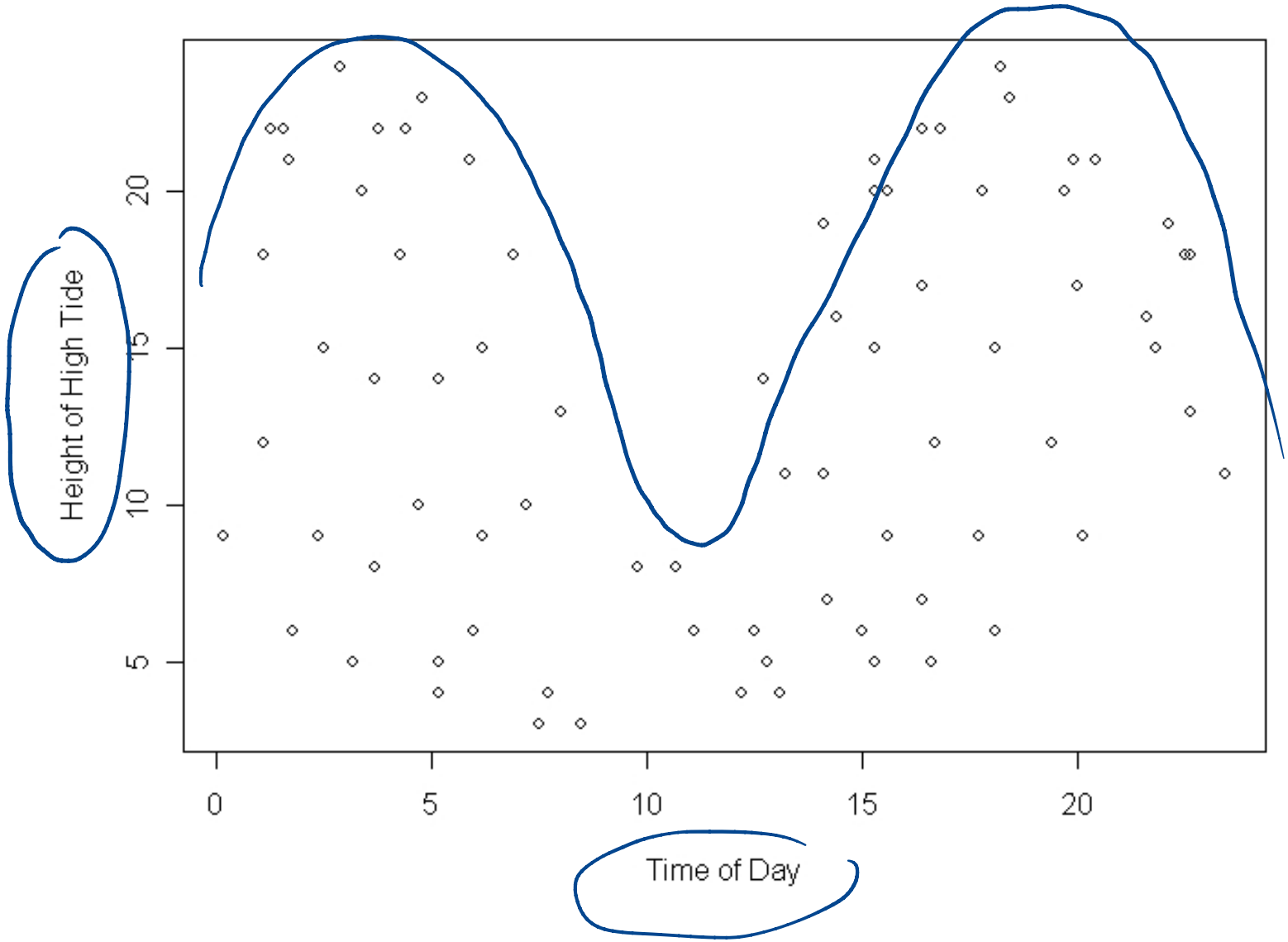
$(y_1, ..., y_n)^\top$ has independent entries, each with **distribution $F(y; \theta_i)$ of the same family but with different parameters**.

- Each observation was generated under **slightly different experimental conditions**. They depend in a similar way on different θ_i.
- These θ_i correspond to different experimental conditions, say x_i.
- Each x_i is called a covariate/feature, and is an input that the experimenter can vary. They are **known**. The index i reminds us that it corresponds to the ith observation y_i.
- Usually θ_i is postulated to have a special relationship to x_i (through g), for example $\theta_i = \exp\{\alpha + \beta x_i\}$, for (α, β) uknown parameters.
- The point here is to understand the effect of varying the covariate/feature on the distribution of the observable.

How to model the height of Honolulu tides throughout the day - Histogram



Height of Honolulu tides as function of the time of day



A **bewildering variety** of models can be captured by the general specificaiton

$$y_i \overset{\text{independent}}{\sim} \text{Distribution}\{\underbrace{g(x_i)}_{=\theta_i}\}, \qquad i = 1, ..., n.$$

x_i 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, Laplace, Bernoulli, 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, neural net...

Table: A coarse classification of regression models we will consider

Distribution / Function g	$g(x_i^\top) = x_i^\top \beta$	g **nonparametric**
Gaussian	Linear Regression	Smoothing
Exponential Family	GLM	GAM

GLM: Generalized Linear Model and GAM: Generalized Additive Model

We start with a very standard model: Linear Regression with $y|x$ being Gaussian.

Fundamental case: Gaussian linear regression

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

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

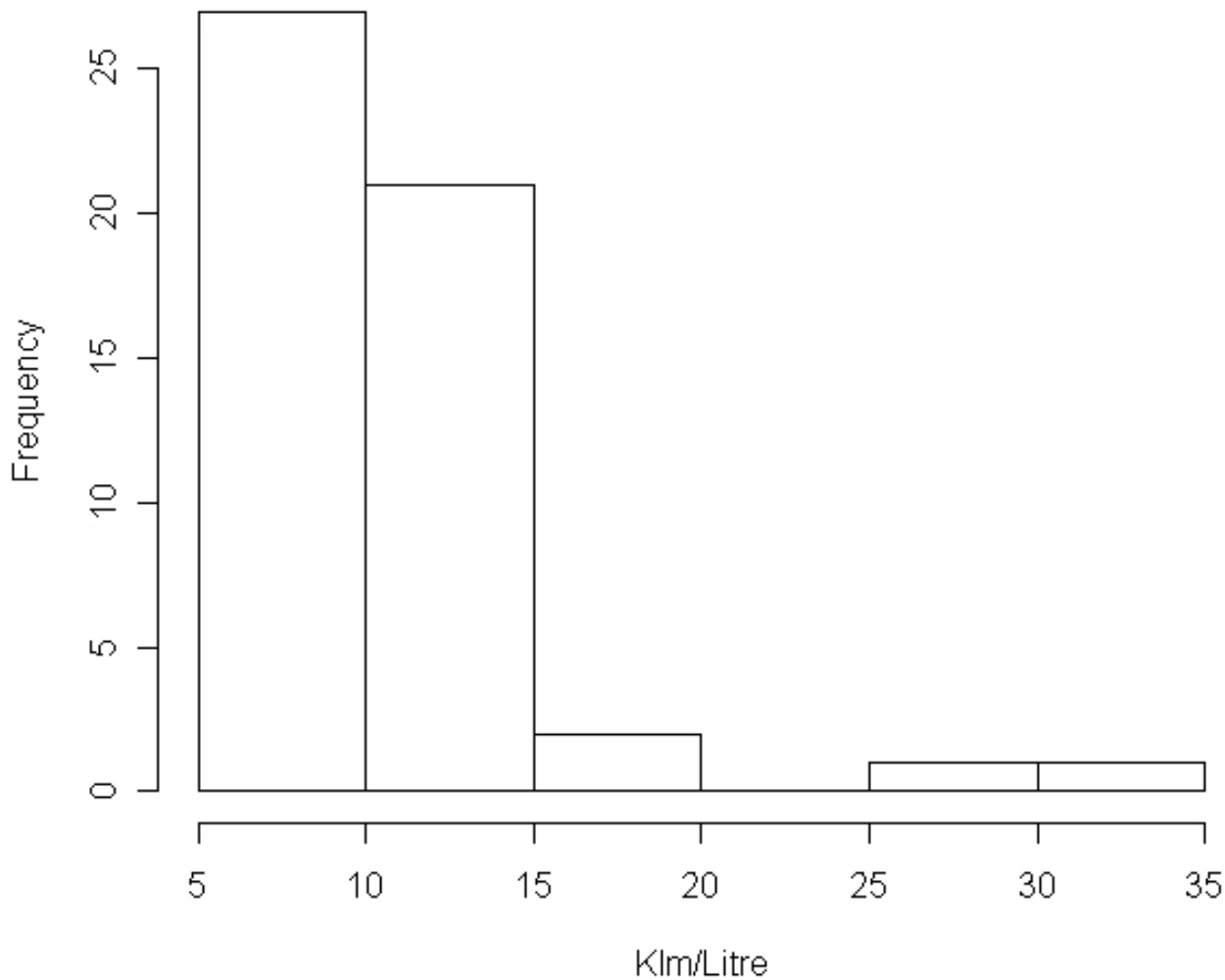
The second version is useful for mathematical work, but is puzzling statistically, since we don't observe ϵ.

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

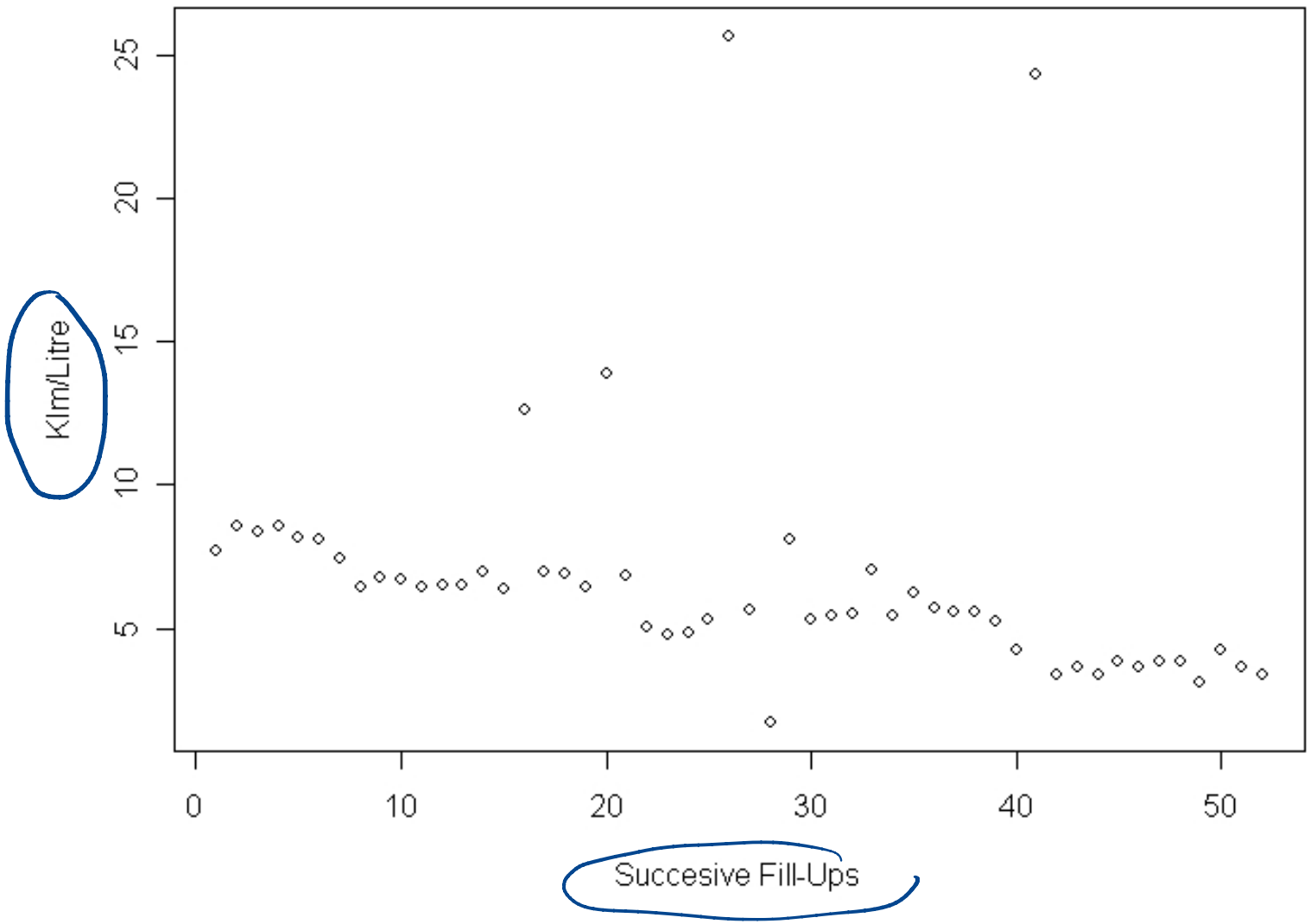
$$y \mid x \sim \mathcal{N}(\beta_0 + \beta^\top x, \sigma^2)$$
$$\Updownarrow$$
$$y = \beta_0 + \beta^\top x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

How to model my van's consumption of gas

Histogram of consumption of gas (km/L)



Gas consumption as function of successive fill-ups



The tools of the trade . . .

Start from **Gaussian linear regression** then gradually generalise . . .

Obviously: important features of Gaussian linear model are

- Gaussian distribution
- Linearity

These two **combine well** and give **geometric insights** to solve the estimation problem related to some **probabilistic linear algebra**. . .

- Subpsaces and projection matrices
- Multivariate Gaussian Distribution
- Optimal dimension reduction
- Random quadratic forms

To sum up

- **Regression**: (statistics) a measure of the relation between the mean value of
  - one variable (e.g., output), denoted y (the **response variable**) and
  - corresponding values of other variables (e.g., time and cost), denoted x (**explanatory variables**).
- The explanatory variables are also called **covariates** or **features** (ML).
- We avoid the terms **dependent variable** (y) and **independent variable** (x) used in older books.
- Questions we try and answer:
  - (**description/explanation**) how does y depend on x? How much of the variation of y is due to x? Do I need all of x to explain the variation in y?
  - (**prediction**) what will y be if $x = x_+$?
  - (**causation**) if I change x, what will happen to y?
- The causation question presupposes that we can change (some of) x, which is not always true.

The Linear Model - First Results and Geometric Interpretations

Linear model

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

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

where

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

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

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

and we write

$$y = X\beta.$$

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

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

Notation

- Vectors are column vectors
- We write $X_{n\times p}$ to give the dimensions of a matrix or vector
- a^{T} (row vector) is the transpose of a (column vector)
- $j \in \{1, \ldots, n\}$ (or sometimes i) indexes the rows of y (cases/examples)
- x_j^{T} is the jth row of X
- $r, s, t, \ldots \in \{1, \ldots, p\}$ indexes the columns of X (covariates/features)
- Roman letters (y, X, z, ...) denote observed quantities, and may be the realisations of random variables
- Greek letters ($\beta, \gamma, \theta, \sigma, \ldots$) denote unknown (often vector) parameters of models
- $\widehat{\beta}$ denotes an estimate of β
- α denotes the level of significance tests and confidence intervals
- If Q is scalar (or a row vector) and β is a vector, then $\partial Q/\partial\beta$ denotes the vector (or matrix) the same shape as β with elements $\partial Q/\partial\beta_r$.
- If Q is scalar and β, γ are vectors, then $\partial^2 Q/\partial\beta\partial\gamma^{\mathrm{T}}$ denotes the matrix with (r, s) element $\partial^2 Q/\partial\beta_r\partial\gamma_s$.
- $u \perp v$ means that the vectors u and v are orthogonal (i.e., $u^{\mathrm{T}}v = 0$); ditto for matrices.
- $Y \perp\!\!\!\perp Z$ means that the random variables Y and Z are independent.

Useful matrix decompositions

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

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

where

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

- The SVD implies that the ranks of X, $X^{\mathrm{T}}X$ and XX^{T} are equal and at most m.

- **Spectral theorem**: any real symmetric matrix H can be written as

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

where

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

Least squares fit

- Assume that

$$y = X\beta$$

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

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

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

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

Optimal regressor

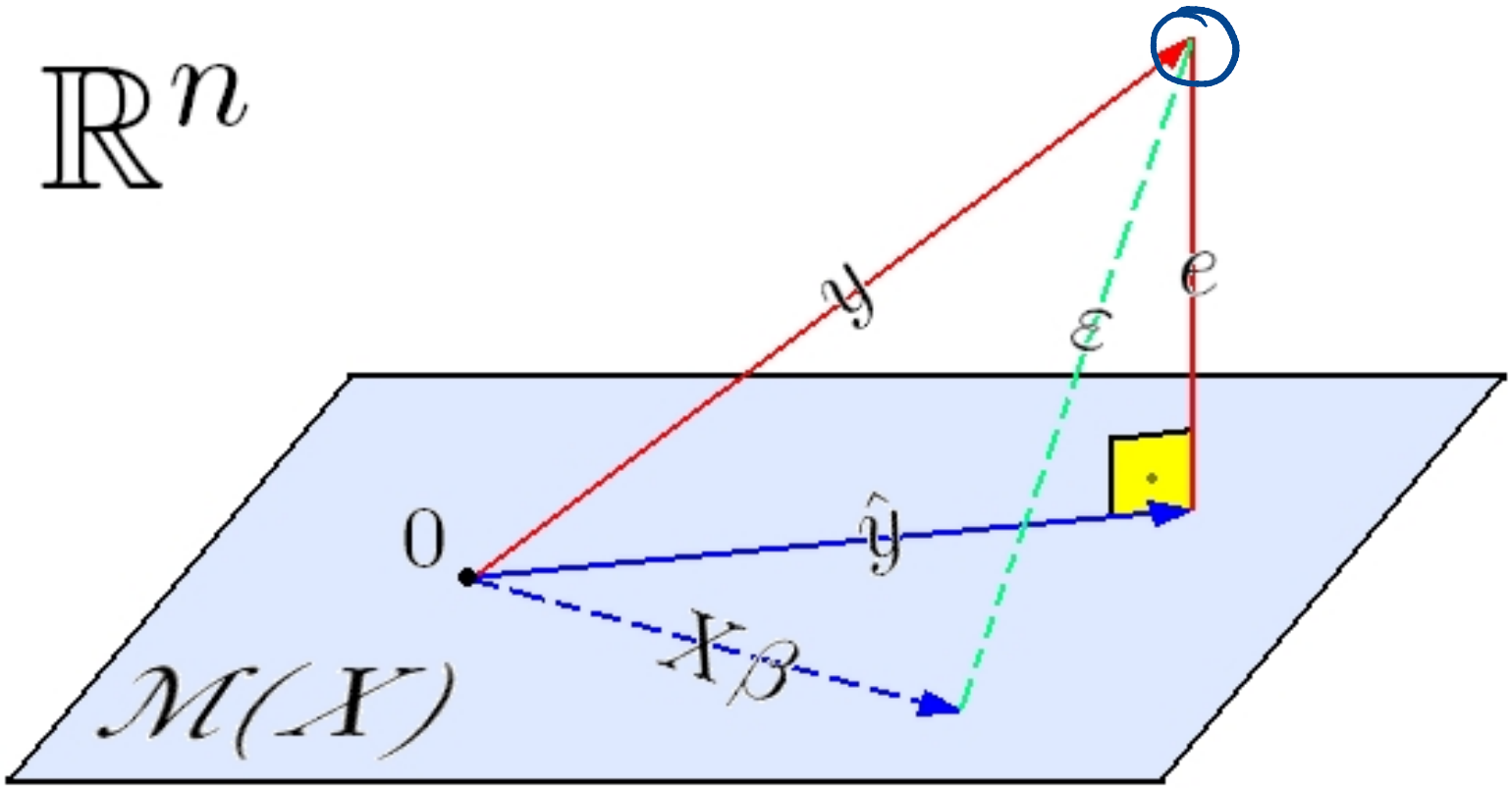
Lemma 1

When X has rank p and $n \geq p$ then $\widehat{y} = X\widehat{\beta} = Hy$, where

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

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

The (column) geometry of least squares



Proof

$Q = \|y - X\beta\|^2 = (y - X\beta)^T(y - X\beta) = y^Ty - 2y^TX\beta + \beta^TX^TX\beta$

$\frac{\partial Q}{\partial \beta} = -2y^TX + 2X^TX\beta \qquad \frac{\partial^2 Q}{\partial \beta \partial \beta^T} = 2X^TX$

$\frac{\partial Q}{\partial \beta} = 0$ if $X^TX\beta = y^TX = X^Ty$

if $(X^TX)^{-1}(X^TX)\beta = (X^TX)^{-1}X^Ty$

- If X has rank p, so too does the $p \times p$ matrix $X^{\mathrm{T}}X$, which is therefore invertible.
- The sum of squares

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

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

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

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

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

say. The matrix $X^{\mathrm{T}}X$ is positive definite, so $(y - X\beta)^{\mathrm{T}}(y - X\beta)$ is minimised at $\widehat{\beta}$.

Note to Lemma 1

- The $n \times n$ 'hat matrix' H (which 'puts a hat' on y) satisfies $H^{\mathrm{T}} = H$, $H^2 = H$, so it is symmetric and idempotent, i.e., its eigenvalues equal 0 or 1, and their multiplicities must be $n - p$ and p, as its rank is p.

  H is the matrix that projects $\mathbb{R}^n$ orthogonally onto the span of the columns of X, $\mathrm{span}(X)$**.**

- The inner product between $\widehat{y}$ and $y - \widehat{y}$ equals zero, because $\widehat{y} = Hy$, $y - \widehat{y} = (I - H)y$, and $\widehat{y}^{\mathrm{T}}(y - \widehat{y}) = y^{\mathrm{T}} H^{\mathrm{T}}(I - H)y = y^{\mathrm{T}}(H - H)y = 0$.

  Hence $\widehat{y}$ and $y - \widehat{y}$ are orthogonal.

- Clearly $HX = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X = X$, so $H(X\beta) = X\beta$ for any $\beta \in \mathbb{R}^p$, i.e., a vector in $\mathrm{span}(X)$ is left unchanged by multiplication by H.

Analysis of variance I

Lemma 2

Let $X_{n\times p} = (X_0, X_1, \ldots, X_R)$ have rank p, where $p \leq n$, and let H_r denote the projection matrices formed using $(X_0, \ldots, X_r)$, for $r = 0, \ldots, R$; hence $H_R = H$.

Define $P_r = H_r - H_{r-1}$ for $r = 1, \ldots, R$ and $P_{R+1} = I - H$.

Then:

(i) $H_r H_s = H_r$ *whenever* $r \leq s$,

(ii) $H_0 P_r = 0$ *for any* r,

(iii) *the matrices* P_r *are symmetric and idempotent, with* $P_r P_s = 0$ *when* $r \neq s$.

Rephrasing the question: Gaussian linear model

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

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

Interpretation: $\widehat{y} = X\widehat{\beta} = Hy$ is the projection of y into the column space of X, $\mathrm{span}(X)$. 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 $\beta_2 = 0$?

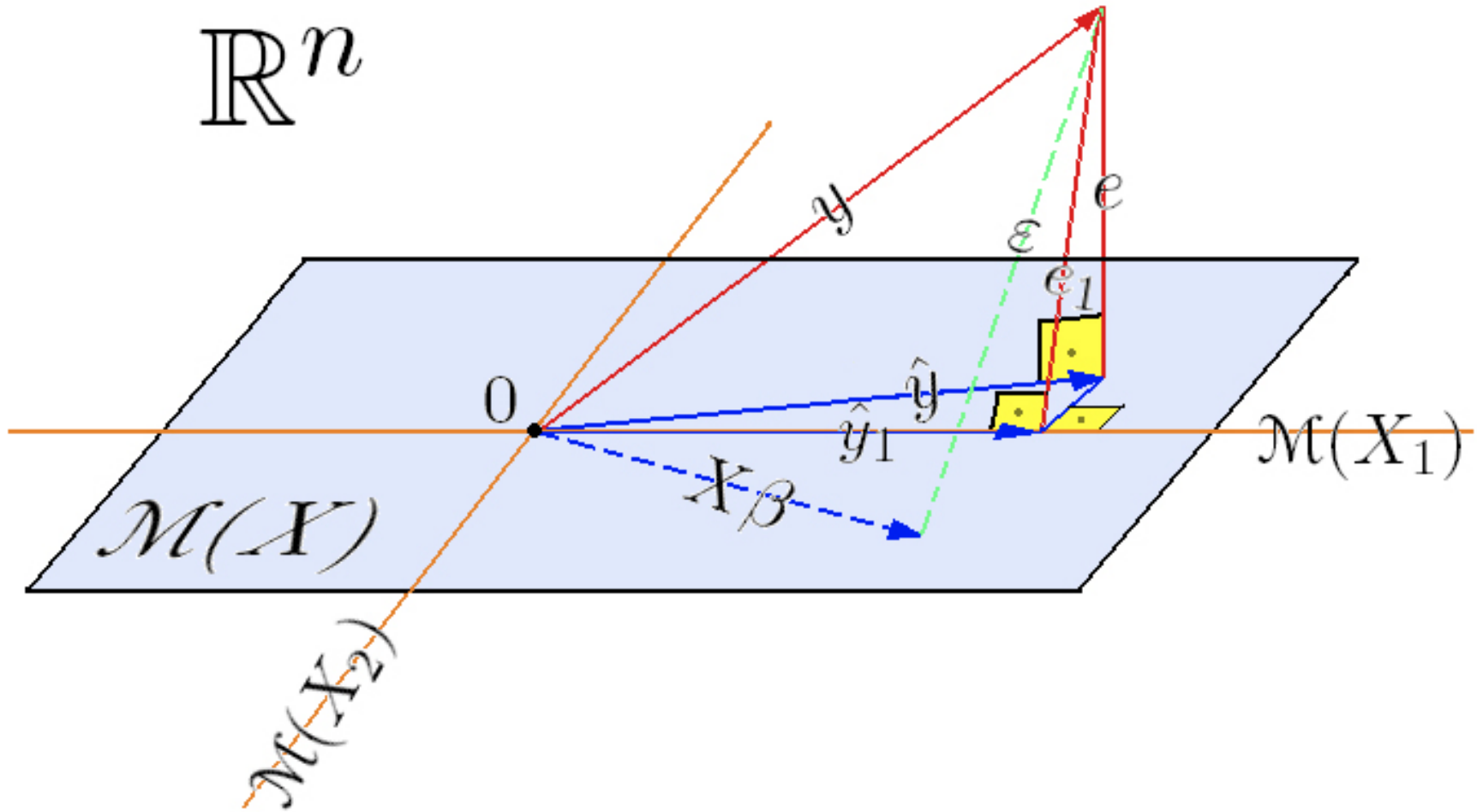


Figure: Geometry Revisited

Proof

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

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

(ii) For $r = 1, \ldots, R$, (i) yields $H_0 P_r = H_0 H_r - H_0 H_{r-1} = H_0 - H_0 = 0$, and $H_0 P_{R+1} = H_0 (I - H_R) = 0$.

(iii) The matrices $P_1, \ldots, P_R$ are symmetric because

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

and idempotent because (i) gives

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

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

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

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

General decomposition

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

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

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

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

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

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

Analysis of variance II

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

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

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

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

  decomposes ('analyses') the variability of y around its average $\overline{y}$ into

  - the contributions $\|\widehat{y}_r - \widehat{y}_{r-1}\|^2$ due to adding the columns of X_r to $X_0, \ldots, X_{r-1}$,
  - the **residual sum of squares** $\|y - \widehat{y}\|^2$ left after fitting $X = (X_0, \ldots, X_R)$.

- Large $\|\widehat{y}_r - \widehat{y}_{r-1}\|^2$ implies that X_r explains a lot of the variation of y even after allowing for that explained by $X_0, \ldots, X_{r-1}$.

- The
  - **degrees of freedom** of a fit is the rank ν_r of the corresponding H_r, and the
  - **residual degrees of freedom** is $n - \nu_R = n - p$.

Terms

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

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

  where the matrices $X_1, \ldots, X_R$, the **terms**, are successively included.

- The baseline model with only 1_n has fitted value and residual vector

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

- Starting from the baseline we ask which terms lead to large reductions in the residual sum of squares, i.e., best explain the variation of y.

- The successive residual degrees of freedom, i.e., the ranks of the matrices $I - H_r$, are

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

  full model.

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

Model formulae

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

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

  where

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

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

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

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

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

ANOVA

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

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

$$\mathbf{MS}_r = \|y - \widehat{y}_r\|^2.$$

- The 'mean square' for term X_r,

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

  is the average reduction in $\mathbf{MS}_r$ per degree of freedom when X_r is added to the model.

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

ANOVA table

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

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

Coefficient of determination

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

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

  where H_0 and H are the hat matrices for regression on 1_n and X, and $1_n \in \mathrm{span}(X)$.

- $R^2 \in [0, 1]$ is the squared empirical correlation between y and $\widehat{y}$, so $R^2 \approx 1$ implies that most of the variation in y is explained by $\widehat{y}$.

- There is a geometric interpretation, as the terms on the right of

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

  are orthogonal (check this).

- Adding columns to X must increase R^2, unlike the **adjusted** R^2,

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

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

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

Comments

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