

1.1 Simple linear regression (repetition)

$$\begin{aligned} \text{Model : } Y &= \beta_0 + \beta_1 x + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned} \quad \left. \vphantom{\begin{aligned} Y &= \beta_0 + \beta_1 x + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned}} \right\} (*)$$

[2min discussion : describe all the components of the model]

Y : the (random) response variable
dependent variable, output, target
→ the phenomenon that we want to investigate.

X : a covariate (explanatory var, predictor, input, feature)
→ a variable that we can use to describe or predict the response.

$\beta_0 + \beta_1 x$: regression line $E(Y|x) = \beta_0 + \beta_1 x$
 β_0 : intercept, β_1 : slope
 β_0, β_1 : coefficients

ε : error / noise represents deviations of the response from the regression line
— measurement error
— lack of information

Since $\varepsilon \sim N(0, \sigma^2)$ and $Y = \beta_0 + \beta_1 X + \varepsilon$

then $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

mean of Y
depends on x

variation around
the regression line is
not dependent on x .

[2 min discussion: which datasets
do you think have been generated from
a model such as (*)?]

A $\checkmark$ B $\checkmark$ C $\checkmark$ D $\times$ $\leftarrow$ variance depends on x .

Typical problem:

we have n pairwise observations

$(x_1, y_1), \dots, (x_n, y_n)$ and want to estimate
the parameters $\beta_0, \beta_1, \sigma^2$ of the model (*).

Where do such data come from?

Data collection methods

(i) Designed experiment

Decide x -values : x_1 x_2 $\dots$ x_n
e.g. dose of medication
 $\downarrow$ $\downarrow$ $\downarrow$
Measure resulting response: y_1 y_2 y_n
e.g. blood pressure

(ii) observational study

Sample n subjects from a population

e.g. Norwegian women 60-70 years

observe (x_i, y_i) for each person i .

physical activity blood pressure.

both response and covariate random,
but we model the response y conditional on x .

With n observations we assume n models

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, 1)$$
$$i = 1, \dots, n$$

We also assume $\varepsilon_i, \varepsilon_j$ independent

Collect n models into a
matrix formulation

$$\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{random response vector}$$

$$\vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ random error vector}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ design matrix}$$

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ vector of coefficients}$$

$$\text{Model: } \vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

random
vectors
next week.

$$\text{observation: } (\vec{y}, X)$$

1.2 Multiple linear regression

For $i=1, \dots, n$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

as before

$k+1$ coefficients $\beta_0, \beta_1, \dots, \beta_k$

k covariates $x_1, x_2, \dots, x_k$.

Matrix formulation

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

↑
design matrix
 $n \times (k+1)$

← $(k+1)$ -vector

require: $n > (k+1)$ more observations than covariates -

and $\text{rank}(X) = k+1$ "full rank"

i.e. one covariate can not be a sum or linear combination of the other covariates -

$X^T X$ pos.
def. and
invertible

Note: we can have models of the type

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 \log(x_2) + \varepsilon$$

As before, we observe $\vec{y}$, and using $\vec{y}$ and X we will estimate $\vec{\beta}$.

(PART 2)

1.3 Analysis of variance (ANOVA)

↳ Compare the response $\vec{Y}$ between groups

i.e. 3 different treatments,

μ_j mean of $\vec{Y}$ under treatment $j, j=1,2,3$

Model $Y_j = \mu_j + \varepsilon$

or $Y_j = \mu + \alpha_j + \varepsilon$

↑
overall
mean

↖ deviation
from overall
mean due to
treatment j

Hypothesis : treatment effects (α_j -s)
differ.

(special case of mult.lin. regression)

Part 2 : Estimate models from data,
distn of estimators, hypothesis test,
checking model assumptions etc.

2.13 Idempotent matrices

Motivation: We will consider 'quadratic forms'

$$\begin{matrix} \vec{y}^T A \vec{y} \\ \uparrow \quad \nearrow \\ \text{random} \quad \text{vector} \end{matrix} \quad \text{where } A \text{ is an idempotent matrix.}$$

→ need to know properties of such matrices

Def: A square matrix A is idempotent if $AA = A$

our idempotent matrices will typically also be symmetric.

Thm 2.13 a) The only nonsingular (invertible) idempotent matrix is the identity matrix I .

Proof: If A idempotent and invertible

then $AA = A$ and A^{-1} exists, $A^{-1}A = I$.

$$AA = A$$

mult both sides by A^{-1} .

$$\underbrace{A^{-1}AA}_{I} = \underbrace{A^{-1}A}_{I} \rightarrow A = I.$$

Thm 2.13 b) If A is symmetric and idempotent and singular (not invertible, i.e. not I), then A is positive semidefinite.

$$\vec{x}^T A \vec{x} \geq 0 \text{ for all } \vec{x} \in \mathbb{R}^n$$

→ our quadratic forms will not take negative values.

Proof sketch : A symmetric $\rightarrow A^T = A$

A idempotent $\rightarrow A A = A$

$$A = A A = \underbrace{A^T A}$$

pos. semidef. by thm 2.6 d).

If time:

Note : For scalar a , $a^2 = a \rightarrow a(a-1) = 0$
 $\rightarrow a = 0$ or $a = 1$

For matrix A , $A A = A$ implies that eigenvalues are either ± 1 or 0 .

Thm 2.13 c) If A is an $n \times n$ symmetric idempotent matrix of rank r , then A has r eigenvalues equal to ± 1 and $n-r$ eigenvalues equal to 0 .

Thm 2.13 d) If A is symmetric and idempotent then

$$\text{rank}(A) = \text{trace}(A)$$

$\uparrow$
linearly independent
columns

$\uparrow$
sum of diagonal
elements,
sum of eigenvalues.

$\hookrightarrow$ Useful because we will need $\text{rank}(A)$ for A in $\vec{y}^T A \vec{y}$.

Note also thm 2.11 (i)

If A and B $n \times n$ then

$$\text{trace}(A \pm B) = \text{trace}(A) \pm \text{trace}(B)$$

Thm 2.13 e) If A is idempotent

(i) $I - A$ is idempotent

Proof: (try first)

$$\begin{aligned} (I - A)(I - A) &= (I - A)I - (I - A)A \\ &= I - A - A + A = I - A. \end{aligned}$$

A special matrix: (Chap 2.1)

$$\mathbf{J} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \quad n \times n \text{ matrix of 1s.}$$

$$\vec{j} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \vec{j}^T \vec{j} = n \quad \vec{j} \vec{j}^T = \mathbf{J}$$

For a vector of observations $\vec{y}$,

$$\vec{j}^T \vec{y} = \sum_{i=1}^n y_i$$

$$\frac{1}{n} \vec{j}^T \vec{y} = \bar{y} \quad (\text{average of all observations})$$

Note: $\frac{1}{n} \mathbf{J}$ is idempotent

$$\frac{1}{n} \mathbf{J} \cdot \frac{1}{n} \mathbf{J} = \frac{1}{n^2} \mathbf{J} \cdot \mathbf{J} = \frac{1}{n^2} \begin{bmatrix} n & \cdots & n \\ \vdots & & \vdots \\ n & \cdots & n \end{bmatrix} = \frac{1}{n} \mathbf{J}$$

$$\text{rank} \left(\frac{1}{n} \mathbf{J} \right) = 1 \quad \text{trace} \left(\frac{1}{n} \mathbf{J} \right) = 1.$$

By thm 2.13 e) $\mathbf{I} - \frac{1}{n} \mathbf{J}$ is also idempotent

Discuss

1) What is $(\mathbf{I} - \frac{1}{n} \mathbf{J}) \vec{y}$

2) What is the rank of $(\mathbf{I} - \frac{1}{n} \mathbf{J})$?

Ans: 1) $\begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$ 2) $\text{trace}(I) - \text{trace}\left(\frac{1}{n}J\right) = n - 1$

3) What is $\vec{y}^T (I - \frac{1}{n}J) \vec{y}$?

Ans: $\vec{y}^T (I - \frac{1}{n}J) (I - \frac{1}{n}J) \vec{y}$
Symmetric
 $((I - \frac{1}{n}J) \vec{y})^T (I - \frac{1}{n}J) \vec{y}$
 $= \sum_{i=1}^n (y_i - \bar{y})^2$

Recall estimator for σ^2 using y_i from $N(\mu, \sigma^2)$ -distr:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

Now: $s^2 = \frac{\vec{y}^T (I - \frac{1}{n}J) \vec{y}}{n-1}$

← idempotent

← quadratic form

← rank $(I - \frac{1}{n}J)$.