Statistical Methods for Economists

Lecture 4

Multiple Linear Regression

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistical Methods for Economists
INM/BASTE

Outline of the lecture

- Introduction: Simple Linear Regression & Least Squares Method
- Multiple Linear Regression: Introduction
- Multiple Linear Regression: Summary & Background
- The Classical Assumptions
- The Coefficient of Determination (R^2)
- Further Theorems, Tests of Hypotheses and Confidence Intervals
- Two-sample *t*-test for the difference of the population means // $\sigma_X=\sigma_Y$
- Simple linear regression without the intercept term

Introduction

- Simple Linear Regression
- Motivation
- Example
- Least Squares Method
- Generalization
- Multiple Linear Regression: Introduction
- Multiple Linear Regression: Notation

Simple Linear Regression: Motivation

Motivation:

Assume a dataset $(y_i, x_{i1})_{i=1}^{n}$ of n statistical units, i.e. we are given n pairs (y_1, x_{11}), (y_2, x_{21}), …, (y_n, x_{n1}) of quantitative variables $(x_{i1}, y_i \in \mathbb{R})$, such as

- x_{i1} = investments and y_i = the resulting revenues
- x_{i1} = particular times and y_i = the price of a stock at the given time
- x_{i1} = the quantity of some goods supplied to a market and y_i = the resulting unit price for the goods
- etc.

Simple Linear Regression: Motivation

Given the n pairs $(y_1, x_{11}),\ (y_2, x_{21}),\ \ldots,\ (y_n, x_{n1})$ of the measurements, we assume that there is a simple linear relationship between the values of X_1 and Y of the form

$$Y \approx \beta_0 + \beta_1 X_1 \qquad \text{for some} \quad \beta_0, \beta_1 \in \mathbb{R}$$

or rather

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad \text{for some} \quad \beta_0, \beta_1 \in \mathbb{R}$$

where ε is a random deviation.

We do not know the parameters β_0 and β_1, however…

Simple Linear Regression: Motivation

Based on the n pairs (y_1, x_{11}), (y_2, x_{21}), …, (y_n, x_{n1}) of the measurements, it is our purpose to find

	the estimates	b_0	and	b_1
of				
	the unknown	β_0	and	β_1

The estimates b_0 and b_1 are also denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, sometimes, i.e. the estimates are

$$b_0 = \hat{\beta}_0 \qquad \text{and} \qquad b_1 = \hat{\beta}_1$$

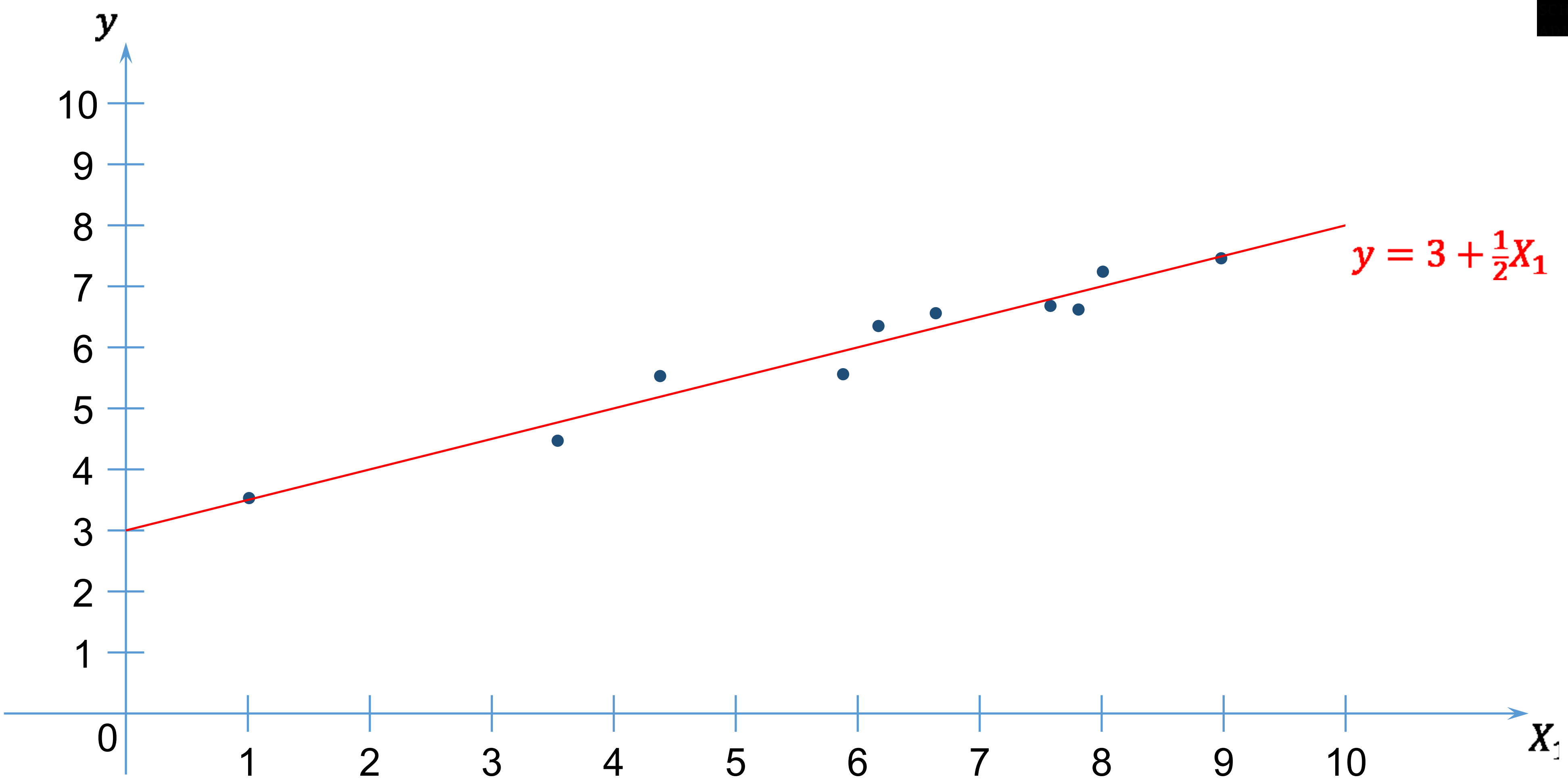
Simple Linear Regression: Example

We have got a sample of $n = 10$ observations:

i	x_{i1}	y_i
□ 1	8.01	7.24
□ 2	7.81	6.62
□ 3	4.38	5.53
□ 4	3.54	4.47
□ 5	6.17	6.35
□ 6	6.64	6.56
□ 7	7.58	6.68
□ 8	8.98	7.46
□ 9	1.01	3.53
10	5.88	5.56

E.g.: x_{i1} = temperature & y_i = the length of a metal rod

Simple Linear Regression: Example



Simple Linear Regression: Least Squares Method

We have got the n pairs (y_1, x_{11}), (y_2, x_{21}), …, (y_n, x_{n1}) of the observations.

For any $b_0, b_1 \in \mathbb{R}$, **the** i-th **estimated value is**

$$\hat{y}_i = b_0 + b_1 x_{i1} \qquad \text{for} \quad i = 1, 2, \dots, n$$

The i-th **residual is the difference**

$$\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

The **residual sum of squares** is

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1})^2$$

Simple Linear Regression: Least Squares Method

Given the n pairs (y_1, x_{11}), (y_2, x_{21}), …, (y_n, x_{n1}) of the observations,

find $b_0, b_1 \in \mathbb{R}$ so that the residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n} (b_0 + b_1 x_{i1} - y_i)^2 \quad \longrightarrow \quad \min$$

is minimized.

The first-order optimality conditions are

$$\frac{\partial \text{RSS}}{\partial b_0} = 0 \qquad \text{and} \qquad \frac{\partial \text{RSS}}{\partial b_1} = 0$$

Simple Linear Regression: Least Squares Method

Given $\text{RSS} = \sum_{i=1}^{n}(b_0 + b_1 x_{i1} - y_i)^2$, we obtain the system of two equations of two unknowns:

$$\frac{\partial \text{RSS}}{\partial b_0} = \sum_{i=1}^{n} 2(b_0 + b_1 x_{i1} - y_i) = 0 \quad \text{and} \quad \frac{\partial \text{RSS}}{\partial b_1} = \sum_{i=1}^{n} 2(b_0 + b_1 x_{i1} - y_i) x_i = 0$$

or

$$n\, b_0 + \sum_{i=1}^{n} x_{i1}\, b_1 = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} x_{i1}\, b_0 + \sum_{i=1}^{n} x_{i1}^2\, b_1 = \sum_{i=1}^{n} x_{i1} y_i$$

the normal equation

Simple Linear Regression: Least Squares Method

Hence,
given the observations $(y_1, x_{11}),\ (y_2, x_{21}),\ \ldots,\ (y_n, x_{n1})$, the estimates are:

$$\hat{\beta}_0 = b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_{i1}\, b_1\right) =$$

$$= \frac{\sum_{i=1}^{n} x_{i1}x_{i1} \sum_{j=1}^{n} y_j - \sum_{i=1}^{n} x_{i1} \sum_{j=1}^{n} x_{j1}y_j}{n\sum_{i=1}^{n} x_{i1}x_{i1} - \sum_{i=1}^{n} x_{i1} \sum_{j=1}^{n} x_{j1}}$$

and

$$\hat{\beta}_1 = b_1 = \frac{n\sum_{i=1}^{n} x_{i1}y_i - \sum_{i=1}^{n} x_{i1} \sum_{j=1}^{n} y_j}{n\sum_{i=1}^{n} x_{i1}x_{i1} - \sum_{i=1}^{n} x_{i1} \sum_{j=1}^{n} x_{j1}}$$

Simple Linear Regression: Generalization

Given the n pairs $(y_1, x_{11}),\ (y_2, x_{21}),\ \ldots,\ (y_n, x_{n1})$ of the measurements,

we have assumed the simple linear relationship of the form

$$Y \approx \beta_0 + \beta_1 X_1 \qquad \text{for some} \quad \beta_0, \beta_1 \in \mathbb{R}$$

The simple linear relationship can be generalized to the form

$$Y \approx \beta_0 X_0 + \beta_1 X_1 \qquad \text{with} \quad X_0 = 1$$

$$\text{for some} \quad \beta_0, \beta_1 \in \mathbb{R}$$

In general, we can have any n triples $(y_1, x_{10}, x_{11}),\ (y_2, x_{20}, x_{21}),\ \ldots,\ (y_n, x_{n0}, x_{n1})$

Simple Linear Regression: Generalization

We shall now study

Multiple Linear Regression

Multiple Linear Regression: Introduction

That is, we are given a dataset $(y_i, x_{i0}, x_{i1}, x_{i2}, \dots, x_{ik})_{i=1}^{n}$ of n statistical units ($(k+2)$-tuples):

$$(y_1, \quad x_{10}, \quad x_{11}, \quad x_{12}, \quad \dots, \quad x_{1k})$$

$$(y_2, \quad x_{20}, \quad x_{21}, \quad x_{22}, \quad \dots, \quad x_{2k})$$

$$\dots$$

$$(y_n, \quad x_{n0}, \quad x_{n1}, \quad x_{n2}, \quad \dots, \quad x_{nk})$$

where $y_i, x_{i0}, x_{i1}, x_{i2}, \dots, x_{ik} \in \mathbb{R}$ for every $i = 1, 2, \dots, n$.

Multiple Linear Regression: Introduction

Given the n $(k+2)$-tuples $(y_i, x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n}$, such as measurements, we assume the multiple linear relationship between the values of $X_0, X_1, \dots, X_k$ and Y of the form

$$Y \approx \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad \text{for some} \quad \beta_0, \beta_1, \dots, \beta_k \in \mathbb{R}$$

or rather

$$Y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \qquad \text{for some} \quad \beta_0, \beta_1, \dots, \beta_k \in \mathbb{R}$$

where ε is a random deviation.

We do not know the parameters $\beta_0, \beta_1, \dots, \beta_k$, however…

Multiple Linear Regression: Notation

We have the dataset of the n $(k+2)$-tuples $(y_i, x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n}$.

The values $(y_i)_{i=1}^{n}$ constitute an n-component column vector $\boldsymbol{y}$, which is an $n \times 1$ matrix, and the $(k+1)$-tuples $(x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n}$ constitute an $n \times (1+k)$ matrix $\boldsymbol{X}$:

$$\boldsymbol{y} = (y_i)_{i=1}^{n} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = (x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

We often have $x_{i0} = 1$ for every $i = 1, 2, \dots, n$,

Multiple Linear Regression: Notation

Assuming the multiple linear relationship between the values of $X_0, X_1, \dots, X_k$ and Y of the form

$$Y \approx X_0\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$$

the unknown parameters $\beta_0, \beta_1, \dots, \beta_k \in \mathbb{R}$ constitute a $(k+1)$-component column vector $\boldsymbol{\beta}$, which is a $(k+1) \times 1$ matrix:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Multiple Linear Regression: Notation

All in all, we have the n equations

$$y_1 = x_{10}\beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1k}\beta_k + \varepsilon_1$$

$$y_2 = x_{20}\beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2k}\beta_k + \varepsilon_2$$

$$\vdots$$

$$y_n = x_{n0}\beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nk}\beta_k + \varepsilon_n$$

where

- the dataset $(y_i, x_{i0}, x_{i1}, \ldots, x_{ik})_{i=1}^{n}$ is given,
- the parameters $\beta_0, \beta_1, \ldots, \beta_k \in \mathbb{R}$ are unknown (to be estimated), and
- the values $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \in \mathbb{R}$ are random deviations (random errors).

Multiple Linear Regression: Notation

The (unknown) random deviations $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \in \mathbb{R}$ constitute an n-component column vector $\boldsymbol{\varepsilon}$, which is an $n \times 1$ matrix:

$$\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^{n} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Moreover, the values $(x_{i0}, x_{i1}, \dots, x_{ik})$ are seen as a $(1+k)$-component row vector $\boldsymbol{x}_i$, which is a $1 \times (1+k)$ matrix:

$$\boldsymbol{x}_i = (x_{i0} \quad x_{i1} \quad \dots \quad x_{ik}) \qquad \text{for} \quad i = 1, 2, \dots, n$$

Multiple Linear Regression: Notation

To sum up, assuming $n \geq 2$ and $k \geq 0$, we have:

$$\boldsymbol{y} = (y_i)_{i=1}^{n} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = (x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n} = \begin{pmatrix} x_{10} & x_{11} & \dots & x_{1k} \\ x_{20} & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^{n} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multiple Linear Regression: Notation

The n equations

$$
\begin{aligned}
y_1 &= x_{10}\beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1k}\beta_k + \varepsilon_1 &= \boldsymbol{x}_1\boldsymbol{\beta} + \varepsilon_1 \\
y_2 &= x_{20}\beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2k}\beta_k + \varepsilon_2 &= \boldsymbol{x}_2\boldsymbol{\beta} + \varepsilon_2 \\
&\vdots \\
y_n &= x_{n0}\beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nk}\beta_k + \varepsilon_n &= \boldsymbol{x}_n\boldsymbol{\beta} + \varepsilon_n
\end{aligned}
$$

can then be written briefly as

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

Random vectors

- Random variable
- Random vector
- Mean value
- Variance-covariance matrix
- Uncorrelated random variables
- Independent random variables

Random variable

Let $(\Omega, \mathcal{F}, P)$ be a **probability space.** That is,

- Ω — the sample space (a non-empty set),
- $\mathcal{F}$ — the event space (a σ-algebra on the sample space Ω)
- P — the probability measure on $(\Omega, \mathcal{F})$.

Recall that a **random variable** is a function

$$X : \Omega \longrightarrow \mathbb{R}$$

which is measurable, i.e. the preimage of any open interval is an event
$(X^{-1}((a,b)) = \{\omega \in \Omega : X(\omega) \in (a,b)\} \in \mathcal{F}$ for every $a, b \in \mathbb{R}$ such that $a < b$).

Random vector

Let $(\Omega, \mathcal{F}, P)$ be the probability space as above, and let n random variables $X_1, X_2, \dots, X_n$ be given. We can then stack the random variables into an n-dimensional **random vector**

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

which is a (measurable) mapping

$$\boldsymbol{X}: \Omega \to \mathbb{R}^n$$

Random vector

__Remark:__ The fact that the mapping $\boldsymbol{X}: \Omega \to \mathbb{R}^n$ is measurable means that the preimage of any open set is an event:

$$\boldsymbol{X}^{-1}(G) = \{\omega \in \Omega : \boldsymbol{X}(\omega) \in G\} \in \mathcal{F} \qquad \text{for every} \quad \text{open } G \subseteq \mathbb{R}^n$$

We assume for simplicity that $\Omega = \mathbb{R}^n$ and that the mapping is the identity:

$$\boldsymbol{X}: \omega \mapsto \omega$$

(the event space $\mathcal{F}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}^n$)

__Remark:__ The mapping $\boldsymbol{X}: \Omega \to \mathbb{R}^n$ is measurable if and only if

Random vector: Expected value

Given the random variables $X_1, X_2, \dots, X_n$,

the **expected value** of the random vector $\boldsymbol{X}$ is:

$$\mathrm{E}[\boldsymbol{X}] = \begin{pmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_n] \end{pmatrix}$$

Random variables: Variance and Covariance

Given the probability space $(\Omega, \mathcal{F}, P)$, let $X: \Omega \to \mathbb{R}$ and $Y: \Omega \to \mathbb{R}$ be some two random variables.

The **variance** of the random variable X is:

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

The **covariance** of the random variables X and Y is:

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

Observation:

Random vector: Variance-covariance matrix

Given the random variables $X_1, X_2, \dots, X_n$,

the **variance-covariance matrix** of the random vector X is:

$$\mathrm{Var}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{cov}(X_1,X_2) & \mathrm{cov}(X_1,X_3) & \dots & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_2,X_1) & \mathrm{Var}(X_2) & \mathrm{cov}(X_2,X_3) & \dots & \mathrm{cov}(X_2,X_n) \\ \mathrm{cov}(X_3,X_1) & \mathrm{cov}(X_3,X_2) & \mathrm{Var}(X_3) & \dots & \mathrm{cov}(X_3,X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \mathrm{cov}(X_n,X_3) & \dots & \mathrm{Var}(X_n) \end{pmatrix}$$

where

$$\mathrm{cov}(X_i,X_j) = \mathrm{E}\left[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])\right] \quad \text{and} \quad \mathrm{Var}(X_i) = \mathrm{cov}(X_i,X_i)$$

Random vector: Uncorrelated random variables

The random variables $X_1, X_2, \dots, X_n$ are (pairwise) **uncorrelated** if and only if

$$\operatorname{cov}(X_i, X_j) = 0 \qquad \text{if} \quad i \neq j \qquad \text{for all} \quad i, j = 1, 2, \dots, n$$

Random vector: Independent events

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A_1, A_2, \dots, A_n \in \mathcal{F}$ be events.

Recall that the events $A_1, A_2, \dots, A_n$ are **mutually independent** if and only if

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \times P(A_2) \times \cdots \times P(A_n)$$

Random vector: Independent random variables

Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space and let $X_1, X_2, \dots, X_n : \Omega \to \mathbb{R}$ be random variables.

The random variables $X_1, X_2, \dots, X_n$ are **mutually independent** if and only if

$$P\begin{pmatrix} \{\omega \in \Omega : a_1 < X_1(\omega) < b_1\} \cap \\ \cap \{\omega \in \Omega : a_2 < X_2(\omega) < b_2\} \cap \\ \dots \\ \cap \{\omega \in \Omega : a_n < X_n(\omega) < b_n\} \end{pmatrix} = \begin{matrix} P\{\omega \in \Omega : a_1 < X_1(\omega) < b_1\} \times \\ \times P\{\omega \in \Omega : a_2 < X_2(\omega) < b_2\} \times \\ \dots \\ \times P\{\omega \in \Omega : a_n < X_n(\omega) < b_n\} \end{matrix}$$

for every $a_1, b_1, a_2, b_2, \dots, a_n, b_n \in \mathbb{R} \cup \{-\infty, +\infty\}$ such that $a_i < b_i$

Random vector: Independent random variables

Theorem: If the random variables

$X_1, X_2, \dots, X_n$ are **mutually independent,** then they are **pairwise uncorrelated.**

Remark: ¡ The converse does not hold true in general !

Remark: The proof of the theorem is easy if the sample space is finite ($\Omega = \{1, 2, \dots, N\}$) or countable ($\Omega = \{1, 2, 3, \dots\}$). The proof is somewhat involved in the general case (requires some knowledge of the theory of the Lebesgue integral, uses limiting steps – Levi's Theorem).

Multivariate normal distribution

Normal distribution

Consider a probability space $(\Omega, \mathcal{F}, P)$ where the sample space $\Omega = \mathbb{R}$, the event space $\mathcal{F}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$, and the probability P is given by its probability density function

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for} \quad x \in \mathbb{R}$$

for some $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$, so that $\sigma^2 > 0$.

That is, the probability is

$$P(A) = \int_A \varphi(x)\,\mathrm{d}x \qquad \text{for any} \quad A \in \mathcal{F}$$

Normal distribution

Given the above probability space $(\Omega, \mathcal{F}, P)$, the probability P being given by its density

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for} \quad x \in \mathbb{R}$$

then the identity random variable $X: \mathbb{R} \to \mathbb{R}$

$$X(x) = x \qquad \text{for} \quad x \in \mathbb{R}$$

follows the Gaussian normal distribution.

We then say that X is a **Gaussian normal random variable** and write

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Normal distribution

<u>Theorem:</u> Let $(\Omega, \mathcal{F}, P)$ be a probability space. (Consider $\Omega = \mathbb{R}^n$ for simplicity.)

If the random variables $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, …, $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$ are **mutually Independent** and normally distributed, then

$$X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(\mu_1 + \mu_2 + \cdots + \mu_n,\ \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$$

that is, their sum is also normally distributed.

Variance-covariance matrix

Theorem: Let $(\Omega, \mathcal{F}, P)$ be a probability space (with $\Omega = \mathbb{R}^n$ for simplicity) and let $X_1, X_2, \dots, X_n: \Omega \to \mathbb{R}$ be any random variables, which are stacked into a random vector $\boldsymbol{X}$. Then its variance-covariance matrix

$$\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X})$$

is symmetric and positively semi-definite.

That is, it holds

$$\boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\Sigma} \qquad \text{and} \qquad \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u} \geq 0 \qquad \text{for every} \quad \boldsymbol{u} \in \mathbb{R}^n$$

Variance-covariance matrix: A decomposition

Theorem:

Let $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ be any symmetric and positively semi-definite matrix, and let

$$k = \operatorname{rank}(\mathbf{\Sigma})$$

Then there exists a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times k}$ such that

$$\mathbf{\Sigma} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$$

Remark: The matrix $\boldsymbol{A}$ can be obtained • either from the spectral decomposition / eigendecomposition of the matrix $\mathbf{\Sigma}$: $\mathbf{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$ where $\boldsymbol{\Lambda}$ is diagonal and $\boldsymbol{Q}$ is orthonormal ($\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{I}$); • or from the Cholesky decomposition: $\mathbf{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ where

Standard multivariate normal distribution (of dim. $k \geq 1$)

Consider a probability space $(\Omega, \mathcal{F}, P)$ where the sample space $\Omega = \mathbb{R}^k$, the event space $\mathcal{F}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}^k$, and the probability P is given by the standardized normal density function

$$\varphi_k(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k}} \mathrm{e}^{-\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{2}} \qquad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^k$$

That is, the probability is

$$P(A) = \int_A \varphi_k(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad \text{for any} \quad A \in \mathcal{F}$$

Standard multivariate normal distribution (of dim. $k \geq 1$)

Given the above probability space $(\Omega, \mathcal{F}, P)$, the probability P being given by its density

$$\varphi_k(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k}} \mathrm{e}^{-\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{2}} \qquad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^k$$

then the identity random vector $\boldsymbol{Z}: \mathbb{R}^k \to \mathbb{R}^k$

$$\boldsymbol{Z}(\boldsymbol{x}) = \boldsymbol{x} \qquad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^k$$

follows the standard Gaussian multivariate normal distribution.

We then say that $\boldsymbol{Z}$ is a **standard multivariate normal random vector** and write

$$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

Multivariate normal distribution

Let a vector $\mu \in \mathbb{R}^n$ (mean values) and a symmetric positively semi-definite matrix $\Sigma \in \mathbb{R}^{n\times n}$ (variance-covariance matrix) of rank k be given. Moreover, let $A \in \mathbb{R}^{n\times k}$ be a matrix such that $\Sigma = AA^\mathrm{T}$. Finally, consider the probability space $(\Omega, \mathcal{F}, P)$ with the sample space $\Omega = \mathbb{R}^k$ and the standard multivariate normal random variable $Z \sim \mathcal{N}(0, I)$.

Then the random vector $X: \mathbb{R}^k \to \mathbb{R}^n$ defined so that

$$X(x) = AZ(x) \qquad \text{for} \quad x \in \mathbb{R}^k$$

follows the standard Gaussian multivariate normal distribution.

We then say that X is a **multivariate normal random vector** and write

Multivariate normal distribution: Density

If the variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is non-singular,

that is $\operatorname{rank}(\boldsymbol{\Sigma}) = k = n$, then the **probability density function**

of the multivariate normal probability distribution

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

is

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}} \mathrm{e}^{\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}} \qquad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^n$$

If the variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is singular, that is $\operatorname{rank}(\boldsymbol{\Sigma}) = k < n$,

then the **probability density function** of the multivariate normal probability

Multivariate normal distribution: Another definition

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\boldsymbol{X}: \Omega \to \mathbb{R}^n$ be a random vector.

Then $\boldsymbol{X}$ follows a multivariate normal distribution, that is $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

for some $\boldsymbol{\mu} \in \mathbb{R}^n$ and for some symmetric and positively semi-definite $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$

if and only if,

for every $\boldsymbol{a} \in \mathbb{R}^n$, the random variable $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$ is normally distributed;

that is, there exist a $\mu_{\boldsymbol{a}} \in \mathbb{R}$ and a non-negative $\sigma_{\boldsymbol{a}}^2 \in \mathbb{R}$ such that

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n \sim \mathcal{N}(\mu_{\boldsymbol{a}}, \sigma_{\boldsymbol{a}}^2) \qquad \text{for every} \quad \boldsymbol{a} \in \mathbb{R}^n$$

Multivariate normal distribution: Linear transformation

Theorem: Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X: \Omega \to \mathbb{R}^n$ be a multivariate normally distributed random vector, that is $X \sim \mathcal{N}(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^n$ and for some symmetric and positively semi-definite $\Sigma \in \mathbb{R}^{n \times n}$.

Then

$$AX \sim \mathcal{N}\left(A\mu, A\Sigma A^{\mathrm{T}}\right) \qquad \text{for any matrix} \quad A \in \mathbb{R}^{m \times n}$$

Multivariate normal distribution: Theorem

<u>**Theorem:**</u> Let random variables $X_1, X_2, \dots, X_n$ be stacked into a multivariate <u>normally distributed random vector</u> $\boldsymbol{X}$, that is $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some $\boldsymbol{\mu} \in \mathbb{R}^n$ and for some symmetric and positively semi-definite $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. Then

the random variables $X_1, X_2, \dots, X_n$ are **mutually independent**

<u>if and only if</u>

the random variables $X_1, X_2, \dots, X_n$ are **pairwise uncorrelated**

(that is, the variance-covariance matrix $\boldsymbol{\Sigma}$ is diagonal).

<u>Remark:</u>

$\Longrightarrow$ holds true in general, see above

Multiple Linear Regression: Summary & Background

- Summary
- Terminology
- Assumptions
- Random vectors
- The classical assumptions
- Notation

Multiple Linear Regression: Summary

We have got the sample of the n $(k+2)$-tuples

$$(\, y_i,\ \boldsymbol{x}_i \,) = (\, y_i,\ \ x_{i0},\ x_{i1},\ x_{i2},\ \ldots,\ x_{ik} \,) \qquad \text{for} \quad i = 1, 2, \ldots, n$$

of the observations, where $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^{1\times(1+k)}$ or $x_{i0}, x_{i1}, \ldots, x_{ik} \in \mathbb{R}$ for $i = 1, 2, \ldots, n$.

The sample could have been obtained in either of the following two ways:

(see the next two slides)

Multiple Linear Regression: Summary

First:

- A sample of n statistical units was selected from a larger population.
- Each of the statistical units was measured and we have obtained the pairs $(y_i, \boldsymbol{x}_i)$ for $i = 1, 2, \dots, n$ thus.
- ¡¡¡ The values $\boldsymbol{x}_i \in \mathbb{R}^{1\times(1+k)}$ were measured / are known exactly !!! (That is, the values $\boldsymbol{x}_i$ are non-random.)
- We assume $y_i \approx \boldsymbol{x}_i\boldsymbol{\beta}$ and we have $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i$, where ε_i is a random deviation (error).
- The random deviation is caused by the intrinsic properties of the statistical unit

Multiple Linear Regression: Summary

Second:

— We prepared the values $x_1, x_2, \dots, x_n \in \mathbb{R}^{1\times(1+k)}$ at the beginning.

— ¡¡¡ These values $x_1, x_2, \dots, x_n$ are known exactly therefore !!!

— When making the i-th measurement,

we set up the system (adjust the system's setting to x_i exactly) first and

we measure the value y_i of the dependent variable then.

— The random deviation ε_i here is caused

either by the intrinsic properties of the system (further unknown / "random" /

unconsidered factors),

Multiple Linear Regression: Summary

Remarks:

- In practice, the data may be obtained in either way (first or second).
- In either case (first or second), the independent values $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ are assumed to be known exactly, i.e. without any measurement errors.
- Assuming $y_i \approx \boldsymbol{x}_i \boldsymbol{\beta}$, even the dependent values y_i may be measured exactly, i.e. without any measurement error, the random deviation $\varepsilon_i = y_i - \boldsymbol{x}_i \boldsymbol{\beta}$ being caused by the intrinsic properties (other unknown / “random” / unconsidered factors).
- For the purpose of the mathematical analysis, we assume the second case only.

Multiple Linear Regression: Terminology

Parameters

Regression coefficients

$$Y = X_0\beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \varepsilon$$

Regressand

Predicand

Explained variable

Dependent variable

Endogenous variable

Controlled variable

Response

Outcome

Predicted variable

Measured variable

Regressors

Predictors

Explanatory variables

Independent variables

Exogenous variables

Control variables

Stimuli

Covariates

Deviation

Error term

Disturbance

Noise

Multiple Linear Regression: Terminology

If $X_0 = 1$:

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k + \varepsilon$$

The **intercept term**

Parameters

Regression coefficients

Regressand	**Regressors**	**Deviation**
Predicand	Predictors	Error term
Explained variable	Explanatory variables	Disturbance
Dependent variable	Independent variables	Noise
Endogenous variable	Exogenous variables	
Controlled variable	Control variables	
Response	Stimuli	
Outcome	Covariates	
Predicted variable		
Measured variable		

Multiple Linear Regression: Assumptions

- The n row vectors $x_1, x_2, \dots, x_n \in \mathbb{R}^{1\times(1+k)}$ are known exactly, fixed, given before the measurements.
- We have n random variables $Y_1, Y_2, \dots, Y_n$
  and n random variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$.
- We assume that the random variables $Y_1, Y_2, \dots, Y_n$ are independent
  and the random variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent.
- Remark:

  It is enough to assume that the random variables $Y_1, Y_2, \dots, Y_n$ are uncorrelated
  and the random variables $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are uncorrelated.

Multiple Linear Regression: Assumptions

- Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space.
- We stack the random variables $Y_1, Y_2, \ldots, Y_n$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ into **random vectors:**

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

which are (measurable) mappings

$$\boldsymbol{Y}: \Omega \to \mathbb{R}^n \quad \text{and} \quad \boldsymbol{\varepsilon}: \Omega \to \mathbb{R}^n$$

Multiple Linear Regression: Random vectors

- <u>We assume for simplicity</u> that $\Omega = \mathbb{R}^n$ and that the mappings are identities:

$$\boldsymbol{Y}: \omega \mapsto \omega \qquad \text{and} \qquad \boldsymbol{\varepsilon}: \omega \mapsto \omega$$

(the event space $\mathcal{F}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}^n$)

- The **expected values** of the random vectors $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are:

$$\mathrm{E}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{E}[Y_1] \\ \mathrm{E}[Y_2] \\ \vdots \\ \mathrm{E}[Y_n] \end{pmatrix} \qquad \text{and} \qquad \mathrm{E}[\boldsymbol{\varepsilon}] = \begin{pmatrix} \mathrm{E}[\varepsilon_1] \\ \mathrm{E}[\varepsilon_2] \\ \vdots \\ \mathrm{E}[\varepsilon_n] \end{pmatrix}$$

Multiple Linear Regression: Random vectors

- The **variance-covariance matrix** of the random vector $\boldsymbol{Y}$ is:

$$\mathrm{Var}(\boldsymbol{Y}) = \begin{pmatrix} \mathrm{Var}(Y_1) & \mathrm{cov}(Y_1, Y_2) & \mathrm{cov}(Y_1, Y_3) & \dots & \mathrm{cov}(Y_1, Y_n) \\ \mathrm{cov}(Y_2, Y_1) & \mathrm{Var}(Y_2) & \mathrm{cov}(Y_2, Y_3) & \dots & \mathrm{cov}(Y_2, Y_n) \\ \mathrm{cov}(Y_3, Y_1) & \mathrm{cov}(Y_3, Y_2) & \mathrm{Var}(Y_3) & \dots & \mathrm{cov}(Y_3, Y_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_n, Y_1) & \mathrm{cov}(Y_n, Y_2) & \mathrm{cov}(Y_n, Y_3) & \dots & \mathrm{Var}(Y_n) \end{pmatrix}$$

Recall that

$$\mathrm{cov}(Y_i, Y_j) = \mathrm{E}\big[(Y_i - \mathrm{E}[Y_i])(Y_j - \mathrm{E}[Y_j])\big] \quad \text{and} \quad \mathrm{Var}(Y_i) = \mathrm{cov}(Y_i, Y_i)$$

Multiple Linear Regression: Random vectors

- **The variance-covariance matrix** of the random vector $\boldsymbol{\varepsilon}$ is:

$$\operatorname{Var}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \operatorname{Var}(\varepsilon_1) & \operatorname{cov}(\varepsilon_1, \varepsilon_2) & \operatorname{cov}(\varepsilon_1, \varepsilon_3) & \dots & \operatorname{cov}(\varepsilon_1, \varepsilon_n) \\ \operatorname{cov}(\varepsilon_2, \varepsilon_1) & \operatorname{Var}(\varepsilon_2) & \operatorname{cov}(\varepsilon_2, \varepsilon_3) & \dots & \operatorname{cov}(\varepsilon_2, \varepsilon_n) \\ \operatorname{cov}(\varepsilon_3, \varepsilon_1) & \operatorname{cov}(\varepsilon_3, \varepsilon_2) & \operatorname{Var}(\varepsilon_3) & \dots & \operatorname{cov}(\varepsilon_3, \varepsilon_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\varepsilon_n, \varepsilon_1) & \operatorname{cov}(\varepsilon_n, \varepsilon_2) & \operatorname{cov}(\varepsilon_n, \varepsilon_3) & \dots & \operatorname{Var}(\varepsilon_n) \end{pmatrix}$$

Recall that

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}[(\varepsilon_i - \mathrm{E}[\varepsilon_i])(\varepsilon_j - \mathrm{E}[\varepsilon_j])] \quad \text{and} \quad \operatorname{Var}(\varepsilon_i) = \operatorname{cov}(\varepsilon_i, \varepsilon_i)$$

Multiple Linear Regression: The Classical Assumptions

- We have the underlying probability space $(\Omega, \mathcal{F}, P)$, with $\Omega = \mathbb{R}^n$ for simplicity.
- Let $\omega \in \Omega$ be the outcome of the random experiment.
- Recalling that $\boldsymbol{X}$ is the $n \times (1+k)$ design matrix, we have

$$\boldsymbol{y} = \boldsymbol{Y}(\omega) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}(\omega)$$

<u>In other words:</u>

- The measured values $y_1, y_2, \dots, y_n$ are the numerical outcomes $Y_1(\omega), Y_2(\omega), \dots, Y_n(\omega)$ of the random experiment.
- The numerical outcomes $Y_1(\omega), Y_2(\omega), \dots, Y_n(\omega)$ are obtained so that the numerical outcomes $\varepsilon_1(\omega), \varepsilon_2(\omega), \dots, \varepsilon_n(\omega)$ of the random experiment

Multiple Linear Regression: The Classical Assumptions

Recall:

¡¡¡ The values of the regressors $x_1, x_2, \dots, x_n$ are **non-random and known** !!!

¡¡¡ The values of the parameters $\beta_0, \beta_1, \dots, \beta_k$ are non-random but **unknown** !!!

We assume that

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \quad \text{for some } \sigma^2 \in \mathbb{R}_0^+$$

where $\boldsymbol{I}$ denotes the $n \times n$ identity matrix

and $\boldsymbol{0}$ denotes the $n \times 1$ zero vector.

Multiple Linear Regression: The Classical Assumptions

The classical assumptions $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ mean that

$$\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{X\beta} \qquad \text{and} \qquad \mathrm{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$$

that is

$$\mathrm{E}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{E}[Y_1] \\ \mathrm{E}[Y_2] \\ \vdots \\ \mathrm{E}[Y_n] \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1\boldsymbol{\beta} \\ \boldsymbol{x}_2\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n\boldsymbol{\beta} \end{pmatrix} \qquad \text{and} \qquad \mathrm{E}[\boldsymbol{\varepsilon}] = \begin{pmatrix} \mathrm{E}[\varepsilon_1] \\ \mathrm{E}[\varepsilon_2] \\ \vdots \\ \mathrm{E}[\varepsilon_n] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

That is

$$\mathrm{E}[Y_i] = \boldsymbol{x}_i\boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{E}[\varepsilon_i] = 0 \qquad \text{for} \quad i = 1, 2, \dots, n$$

Multiple Linear Regression: The Classical Assumptions

The classical assumptions $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ also mean that

$$\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon) = \sigma^2 I = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

That is,

- $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$ for $i = 1, 2, \dots, n$ for some $\sigma^2 \in \mathbb{R}_0^+$ ← **homoscedasticity**, i.e. the variance is the same
- and the random variables $Y_1, Y_2, \dots, Y_n$ or $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are (pairwise) **uncorrelated**.

Multiple Linear Regression: The Classical Assumptions

The classical assumptions $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ finally mean that

- $Y_i \sim \mathcal{N}(\boldsymbol{x}_i \boldsymbol{\beta}, \sigma^2)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, 2, \dots, n$

  where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal (Gaussian) probability distribution with mean μ and variance σ^2 and that

- the random variables $Y_1, Y_2, \dots, Y_n$ or $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are (pairwise) **uncorrelated**.

Remark: It always holds: If the random variables $Y_1, Y_2, \dots, Y_n$ or $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are mutually independent, then they are (pairwise) uncorrelated.

It also holds: If $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, then the random variables

Multiple Linear Regression: The Classical Assumptions

The classical assumption $Y \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I)$ implies that

linearity

$$\mathrm{E}[Y] = X\boldsymbol{\beta}$$

that is

$$\mathrm{E}[Y_i] = \boldsymbol{x}_i\boldsymbol{\beta} = x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k \qquad \text{for} \quad i = 1, 2, \dots, n$$

Multiple Linear Regression: Notation

- The unknown quantities
  - unknown parameters $\beta_0, \beta_1, \dots, \beta_k$
  - unknown (random) deviations $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$

  are denoted by Greek letters

- The estimates of the unknown parameters $\beta_0, \beta_1, \dots, \beta_k$

  are denoted by the respective Latin letters $b_0, b_1, \dots, b_k$

  or by the hat "^" $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$,

  so that $b_0 = \hat{\beta}_0,\ b_1 = \hat{\beta}_1,\ \dots,\ b_k = \hat{\beta}_k$

Multiple Linear Regression: Notation

- The **predicted values** of the dependent variable are denoted by the hat "^":

$$\hat{y}_i = \boldsymbol{x}_i \boldsymbol{b} = x_{i0}b_0 + x_{i1}b_1 + \cdots + x_{ik}b_k \qquad \text{for} \quad i = 1, 2, \dots, n$$

- The unknown (random) deviations $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are denoted by Greek letters.

  The **residuals** are denoted by the respective Latin letters $e_1, e_2, \dots, e_n$

  or by the hat "^" $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n$,

  so that

$$e_1 = \hat{\varepsilon}_1 = y_1 - \hat{y}_1 \qquad e_2 = \hat{\varepsilon}_2 = y_2 - \hat{y}_2 \qquad \dots \qquad e_n = \hat{\varepsilon}_n = y_n - \hat{y}_n$$

Basic Results (Theorems)

- The normal equation
- The predicted values
- Orthogonal projections
- Theorem 1: $\hat{\boldsymbol{y}}$ and $\boldsymbol{e}$ are independent
- Theorem 2: $\hat{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{H})$
- Theorem 3: $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{M})$
- Theorem 4: $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{C})$ if $\text{rank}(\boldsymbol{X}) = k + 1$

Multiple Linear Regression: The Normal Equation

Given the n pairs (y_1, x_1), (y_2, x_2), …, (y_n, x_n) of the observations, the **Residual Sum of Squares** for the estimates $b_0, b_1, \dots, b_k \in \mathbb{R}$ is

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - x_{i0}b_0 - x_{i1}b_1 - \cdots - x_{ik}b_k)^2 \quad \longrightarrow \quad \min$$

It is our purpose to find the estimates $b_0, b_1, \dots, b_k \in \mathbb{R}$ so that the Residual Sum of Squares RSS is <u>minimized</u>. To this end, we let

$$\frac{\partial \mathrm{RSS}}{\partial b_j} = \sum_{i=1}^{n} -2x_{ij}(y_i - x_{i0}b_0 - x_{i1}b_1 - \cdots - x_{ik}b_k) = 0 \qquad \text{for} \quad j = 0, 1, \dots, k$$

Multiple Linear Regression: The Normal Equation

From $\partial RSS / \partial b_j = \sum_{i=1}^{n} -2x_{ij}(y_i - x_{i0}b_0 - x_{i1}b_1 - \cdots - x_{ik}b_k) = 0$, we obtain

the Normal Equation:

$$\sum_{i=1}^{n} x_{ij}(x_{i0}b_0 + x_{i1}b_1 + \cdots + x_{ik}b_k) = \sum_{i=1}^{n} x_{ij}y_i \qquad \text{for} \quad j = 0, 1, \ldots, k$$

Multiple Linear Regression: The Normal Equation

Recall the notation:

$$\boldsymbol{X} = (x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

and

$$\boldsymbol{y} = (y_i)_{i=1}^{n} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Multiple Linear Regression: The Normal Equation

So the normal equation

$$\sum_{i=1}^{n} x_{ij}(x_{i0}b_0 + x_{i1}b_1 + \cdots + x_{ik}b_k) = \sum_{i=1}^{n} x_{ij}y_i \qquad \text{for} \quad j = 0, 1, \ldots, k$$

$$\sum_{i=1}^{n} x_{ij}(\boldsymbol{x}_i\boldsymbol{b}) = \sum_{i=1}^{n} x_{ij}y_i \qquad \text{for} \quad j = 0, 1, \ldots, k$$

can be written as

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Multiple Linear Regression: The Normal Equation

Having the normal equation $(X^{\mathrm{T}}Xb = X^{\mathrm{T}}y)$, where X is an $n \times (1+k)$ matrix, let

$$p = \mathrm{rank}(X)$$

Assume for simplicity that the matrix X is of full rank, that is,

$$p = \mathrm{rank}(X) = k + 1 \leq n$$

NO
(perfect)
multicollinearity

The matrix $X^{\mathrm{T}}X$ is then non-singular; let:

$$C = (X^{\mathrm{T}}X)^{-1}$$

Multiple Linear Regression: The Normal Equation

We have

$$C = \left(X^{\mathrm{T}}X\right)^{-1} = \begin{pmatrix} c_{00} & c_{01} & c_{02} & \cdots & c_{0k} \\ c_{10} & c_{11} & c_{12} & \cdots & c_{1k} \\ c_{20} & c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & c_{k2} & \cdots & c_{kk} \end{pmatrix}$$

The solution to the normal equation

$$X^{\mathrm{T}}Xb = X^{\mathrm{T}}y$$

is then

$$\boxed{b = CX^{\mathrm{T}}y}$$

Multiple Linear Regression: the predicted values

Recall the n equations

$$y_1 = x_{10}\beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1k}\beta_k + \varepsilon_1$$

$$y_2 = x_{20}\beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2k}\beta_k + \varepsilon_2$$

$$\vdots$$

$$y_n = x_{n0}\beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nk}\beta_k + \varepsilon_n$$

where

- the dataset $(y_i, x_{i0}, x_{i1}, \dots, x_{ik})_{i=1}^{n}$ is given,
- the parameters $\beta_0, \beta_1, \dots, \beta_k \in \mathbb{R}$ are unknown (to be estimated), and
- the values $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \in \mathbb{R}$ are random deviations (random errors).

Multiple Linear Regression: the predicted values

Now the **predicted values** are:

$$\hat{y}_1 = x_{10}b_0 + x_{11}b_1 + x_{12}b_2 + \cdots + x_{1k}b_k$$

$$\hat{y}_2 = x_{20}b_0 + x_{21}b_1 + x_{22}b_2 + \cdots + x_{2k}b_k$$

$$\vdots$$

$$\hat{y}_n = x_{n0}b_0 + x_{n1}b_1 + x_{n2}b_2 + \cdots + x_{nk}b_k$$

where

- $b_0 = \hat{\beta}_0, \quad b_1 = \hat{\beta}_1, \quad \ldots, \quad b_k = \hat{\beta}_k$ are the estimates of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k \in \mathbb{R}$,
- $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are the predicted values.

Multiple Linear Regression: the predicted values

Shortly:

— the solution to the normal equation is

$$\boldsymbol{b} = \boldsymbol{C}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

— the predicted values are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Introduce the notation:

$$\boldsymbol{H} = \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$$

The letter “*H*” stands for “hat”:

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$$

Multiple Linear Regression: some properties of *H*

By the construction (by the Least Squares Method: the vector $\hat{\boldsymbol{y}}$ lies in the linear hull of the columns of the matrix $\boldsymbol{X}$ and is as close to $\boldsymbol{y}$ as possible in the Euclidean distance), the matrix

$$\boldsymbol{H} = \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$$

is the **matrix of the orthogonal projection onto the linear subspace**

$$\{\boldsymbol{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$$

(the linear hull of the columns of the matrix $\boldsymbol{X}$)

---

moreover

$$(\boldsymbol{I} - \boldsymbol{H})$$

is the matrix of the orthogonal projection onto the orthogonal complement

Multiple Linear Regression: some properties of *H*

The matrix $\boldsymbol{H} = \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ therefore is:

— idempotent:

$$\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$$

— symmetric:

$$\boldsymbol{H} = \boldsymbol{H}^{\mathrm{T}}$$

— and:

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$$

The residuals are:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

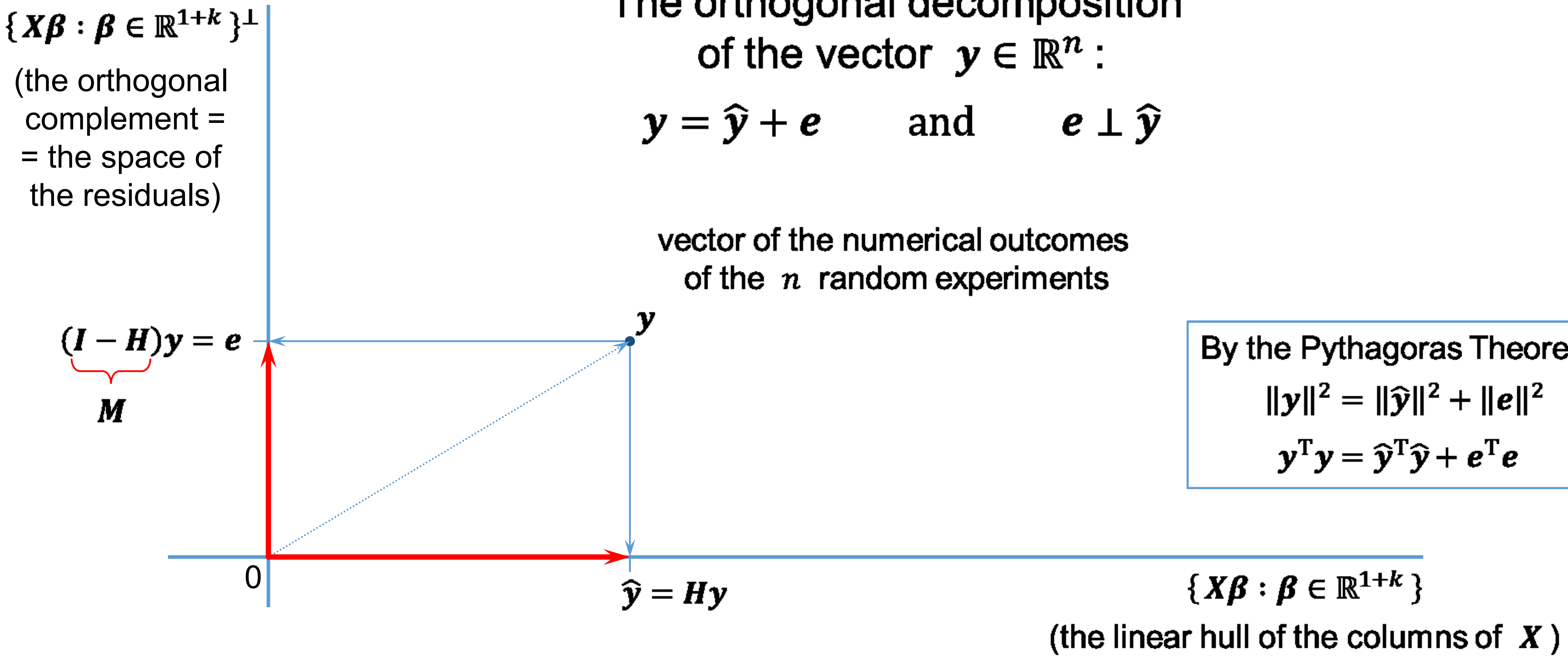
Therefore:

$$\hat{\boldsymbol{y}} \perp \boldsymbol{e}$$

Multiple Linear Regression: some properties of H

The orthogonal decomposition of the vector $\boldsymbol{y} \in \mathbb{R}^n$:

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} \qquad \text{and} \qquad \boldsymbol{e} \perp \hat{\boldsymbol{y}}$$



By the Pythagoras Theorem:

$$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$$

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = \hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}} + \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$

Residual Sum of Squares: $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$

Multiple Linear Regression

Recalling that the regressors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^{1\times(1+k)}$ are given, that we assume

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i \qquad \text{for} \quad i = 1, 2, \dots, n$$

with

$$\mathrm{E}[Y_i] = \boldsymbol{x}_i\boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{Var}(Y_i) = \sigma^2 \qquad \text{for} \quad i = 1, 2, \dots, n$$

or

$$\mathrm{E}[\varepsilon_i] = 0 \qquad \text{and} \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2 \qquad \text{for} \quad i = 1, 2, \dots, n$$

where the random variables $Y_1, Y_2, \dots, Y_n$, or $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, respectively, are independent (or uncorrelated), and that $y_1, y_2, \dots, y_n$ are some observations of the random variables $Y_1, Y_2, \dots, Y_n$, **it follows that all the estimates**

$$b_0, b_1, \dots, b_k \qquad (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) \qquad \hat{y}_1, \hat{y}_2, \dots, \hat{y}_n \qquad \mathrm{RSS} \qquad \text{etc.}$$

Multiple Linear Regression: Theorem 1

Theorem 1: The random vectors $\hat{\boldsymbol{y}}$ and $\boldsymbol{e}$ are independent.

That is,

$$P\begin{pmatrix} \{\omega \in \Omega : \hat{\boldsymbol{y}}(\omega) \in G_{\hat{\boldsymbol{y}}}\} \cap \\ \cap \{\omega \in \Omega : \boldsymbol{e}(\omega) \in G_{\boldsymbol{e}}\} \end{pmatrix} = \begin{matrix} P\{\omega \in \Omega : \hat{\boldsymbol{y}}(\omega) \in G_{\hat{\boldsymbol{y}}}\} \times \\ \times P\{\omega \in \Omega : \boldsymbol{e}(\omega) \in G_{\boldsymbol{e}}\} \end{matrix}$$

for every open set $G_{\hat{\boldsymbol{y}}} \subseteq \mathbb{R}^n$ and for every open set $G_{\boldsymbol{e}} \subseteq \mathbb{R}^n$.

Multiple Linear Regression: Theorem 1: Corollary

Corollary: The random vector $\hat{\boldsymbol{y}}$ and the random variable RSS are independent.

Recall the **Residual Sum of Squares** is $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$

That is,

$$P\begin{pmatrix} \{\omega \in \Omega : \hat{\boldsymbol{y}}(\omega) \in G_{\hat{\boldsymbol{y}}}\} \cap \\ \cap \{\omega \in \Omega : \mathrm{RSS}(\omega) \in G_{\mathrm{RSS}}\} \end{pmatrix} = \begin{matrix} P\{\omega \in \Omega : \hat{\boldsymbol{y}}(\omega) \in G_{\hat{\boldsymbol{y}}}\} \times \\ \times P\{\omega \in \Omega : \mathrm{RSS}(\omega) \in G_{\mathrm{RSS}}\} \end{matrix}$$

for every open set $G_{\hat{\boldsymbol{y}}} \subseteq \mathbb{R}^n$ and for every open set $G_{\mathrm{RSS}} \subseteq \mathbb{R}$

Multiple Linear Regression: Theorem 2

Theorem 2: It holds

$$\hat{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{H})$$

It holds in particular hence that

$$\mathrm{E}[\hat{y}_i] = \mathrm{E}[\boldsymbol{x}_i \boldsymbol{b}] = \boldsymbol{x}_i \boldsymbol{\beta} = \mathrm{E}[Y_i] \qquad \text{for} \quad i = 1, 2, \dots, n$$

and

$$\mathrm{Var}(\hat{y}_i) = \sigma^2 \left(\boldsymbol{XCX}^{\mathrm{T}}\right)_{ii} \qquad \text{for} \quad i = 1, 2, \dots, n$$

$$\mathrm{cov}(\hat{y}_i, \hat{y}_j) = \sigma^2 \left(\boldsymbol{XCX}^{\mathrm{T}}\right)_{ij} \qquad \text{for} \quad i, j = 1, 2, \dots, n$$

Multiple Linear Regression: Theorem 3

Theorem 3: It holds

$$e \sim \mathcal{N}(0, \sigma^2 M)$$

where

$$M = I - H = I - XCX^{\mathrm{T}} = I - X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$$

It holds in particular hence that

$$\mathrm{E}[e_i] = 0 \qquad \text{for} \quad i = 1, 2, \dots, n$$

and

$$\mathrm{Var}(e_i) = \sigma^2 m_{ii} = \sigma^2 - \mathrm{Var}(\hat{y}_i) \qquad \text{for} \quad i = 1, 2, \dots, n$$

$$\mathrm{cov}\left(e_i, e_j\right) = \sigma^2 m_{ij} \qquad \text{for} \quad i, j = 1, 2, \dots, n$$

Multiple Linear Regression: Theorem 4

Theorem 4: If $\text{rank}(\boldsymbol{X}) = k+1$, then

$$\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{C})$$

It holds in particular hence that

$$\text{E}[b_j] = \beta_j \qquad \text{for} \quad j = 0, 1, \ldots, k$$

and

$$\text{Var}(b_j) = \sigma^2 c_{jj} \qquad \text{for} \quad j = 0, 1, \ldots, k$$

$$\text{cov}(b_j, b_i) = \sigma^2 c_{ji} \qquad \text{for} \quad j, i = 0, 1, \ldots, k$$

Residual Sum of Squares, χ^2-test for the variance σ^2, and confidence intervals

- Residual Sum of Squares (RSS)
- Theorem 5: $\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-p}$
- χ^2-test for the variance σ^2
- Confidence intervals

Multiple Linear Regression: Residual Sum of Squares

Recall the **Residual Sum of Squares** is

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i \boldsymbol{b})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

We know that the matrix $\boldsymbol{H}$ is symmetric $(\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{H})$ and idempotent $(\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H})$.

Moreover, we know by the Pythagoras Theorem (see above) that

$$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} =$$

$$= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\hat{\boldsymbol{y}} = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

Multiple Linear Regression: Mean Square Error

Put together, the **Residual Sum of Squares** is

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

Define the **residual variance** or the **Mean Square Error** as

$$s^2 = \frac{\mathrm{RSS}}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\boldsymbol{b})^2}{n-p}$$

where

$$p = \mathrm{rank}(\boldsymbol{X})$$

Multiple Linear Regression: Theorem 5

Theorem 5: It holds

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}$$

where

$$p = \text{rank}(X) \qquad \text{and} \qquad \text{RSS} = \boldsymbol{e}^{\text{T}}\boldsymbol{e} = \boldsymbol{y}^{\text{T}}\boldsymbol{y} - \boldsymbol{y}^{\text{T}}\boldsymbol{X}\boldsymbol{b}$$

Recall that, if $X \sim \chi^2_{n-p}$, then $\text{E}[X] = n - p$.

Therefore:

$$\text{E}[\text{RSS}] = \sigma^2(n-p)$$

$$\text{E}[s^2] = \text{E}\left[\frac{\text{RSS}}{n-p}\right] = \sigma^2$$

Multiple Linear Regression: Theorem 5

Remark: Use Theorem 5 $(\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-p})$

— to obtain an unbiased estimate of the variance:

$$\mathrm{E}[s^2] = \sigma^2 \qquad \text{that is} \qquad s^2 \approx \sigma^2$$

— for a χ^2-test about the variance:

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

— or to establish the confidence intervals for the variance σ^2

Test of hypothesis about the variance σ^2

Remark:

Theorem 5 $(RSS/\sigma^2 \sim \chi^2_{n-p})$ can be used to conduct the χ^2-test for the variance.

- Let $\sigma_0^2 \in \mathbb{R}_0^+$ be a prescribed number.
- Formulate the null hypothesis:

$$H_0: \quad \sigma^2 = \sigma_0^2$$

- Formulate the alternative hypothesis
  - two-sided: $H_1: \quad \sigma^2 \neq \sigma_0^2$
  - one-sided: $H_1: \quad \sigma^2 < \sigma_0^2$
  - one-sided: $H_1: \quad \sigma^2 > \sigma_0^2$

χ^2-test for the variance σ^2

<u>Notation:</u> Let

$$\chi^2_{n-p}(q)$$

denote the **quantile function of Pearson's χ^2-distribution** with $n-p$ d.f., where $p = \text{rank}(\boldsymbol{X})$.

The quantile function $\chi^2_{n-p}(q)$ is the function inverse to the cumulative distribution function $F(x)$ of **Pearson's χ^2-distribution** with $n-p$ degrees of freedom, i.e.

$$\chi^2_{n-p}(q) = F^{-1}(q) \qquad \text{for} \quad q \in (0,1)$$

χ^2-test for the variance σ^2

<u>Notation:</u> Let

$$\chi^2_{n-p}(q)$$

denote the **quantile function of Pearson's χ^2-distribution** with $n - p$ d.f. ,

where $p = \mathrm{rank}(\boldsymbol{X})$.

In other words, if $0 < q < 1$, then $x = \chi^2_{n-p}(q)$ is the unique value such that

$$\int_{-\infty}^{\chi^2_{n-p}(q)} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = q$$

where $f(t)$ is the density of Pearson's χ^2-distribution with $n - p$ d.f.

χ^2-test for the variance σ^2

Having chosen the value $\sigma_0^2 \in \mathbb{R}_0^+$ and assuming the null hypothesis $H_0\colon \sigma^2 = \sigma_0^2$ is true, calculate the statistic

$$X^2 = \frac{\mathrm{RSS}}{\sigma^2} = \frac{\mathrm{RSS}}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma_0^2}$$

χ^2-test for the variance σ^2

The χ^2-test for σ^2 with two-sided alternative hypothesis ($\sigma^2 \neq \sigma_0^2$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical values** are $c = \chi^2_{n-p}\left(\frac{\alpha}{2}\right)$ and $d = \chi^2_{n-p}\left(1 - \frac{\alpha}{2}\right)$
- if $X^2 \in [0, c] \cup [d, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $X^2 \in (c, d)$, then **do not reject** (or fail to reject) the null hypothesis

χ^2-test for the variance σ^2

The χ^2-test for σ^2 with one-sided alternative hypothesis ($\sigma^2 < \sigma_0^2$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.
- the **critical value** is $c = \chi^2_{n-p}(\alpha)$
- if $X^2 \in [0, c]$, **the critical region**, then **reject** the null hypothesis
- if $X^2 \in (c, +\infty)$, then **do not reject** (or fail to reject) the null hypothesis

χ^2-test for the variance σ^2

The χ^2-test for σ^2 with one-sided alternative hypothesis ($\sigma^2 > \sigma_0^2$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical value** is $d = \chi^2_{n-p}(1-\alpha)$
- if $X^2 \in [d, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $X^2 \in [0, d)$, then **do not reject** (or fail to reject) the null hypothesis

Confidence interval for the variance σ^2

Let $x, y \in \mathbb{R}^+$ be any numbers such that $x < y$ and let $F(x)$ be the cumulative distribution function of Pearson's χ^2-distribution with $n - p$ degrees of freedom. Then, by the definition of the cumulative distribution function and by Theorem 5, the probability

$$P\left(x < \frac{\mathrm{RSS}}{\sigma^2} \leq y\right) = F(y) - F(x)$$

Therefore

$$P\left(\frac{\mathrm{RSS}}{y} \leq \sigma^2 < \frac{\mathrm{RSS}}{x}\right) = F(y) - F(x)$$

Confidence interval for the variance σ^2

We have:

$$P\left(\frac{\mathrm{RSS}}{y} \le \sigma^2 < \frac{\mathrm{RSS}}{x}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = \chi^2_{n-p}\left(1 - \frac{\alpha}{2}\right)$ and let $x = \chi^2_{n-p}\left(\frac{\alpha}{2}\right)$. Recall that $\chi^2_{n-p}(q) = F^{-1}(q)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \sigma^2 \in \left[\frac{\mathrm{RSS}}{\chi^2_{n-p}\left(1 - \frac{\alpha}{2}\right)}, \frac{\mathrm{RSS}}{\chi^2_{n-p}\left(1 - \frac{\alpha}{2}\right)}\right]$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the variance σ^2

We have:

$$P\left(\frac{\mathrm{RSS}}{y} \le \sigma^2 < \frac{\mathrm{RSS}}{x}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = \chi^2_{n-p}(1-\alpha)$ and let $x \searrow 0$. Recall that $\chi^2_{n-p}(q) = F^{-1}(q)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \sigma^2 \in \left[\frac{\mathrm{RSS}}{\chi^2_{n-p}(1-\alpha)},\ +\infty\right)$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the variance σ^2

We have:

$$P\left(\frac{\mathrm{RSS}}{y} \leq \sigma^2 < \frac{\mathrm{RSS}}{x}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\ \%$.

Let $y = +\infty$ and let $x = \chi^2_{n-p}(\alpha)$. Recall that $\chi^2_{n-p}(q) = F^{-1}(q)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \sigma^2 \in \left[0,\ \frac{\mathrm{RSS}}{\chi^2_{n-p}(\alpha)}\right]$$

is about $1-\alpha = 95\ \%$.

t-test for a single linear combination of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ — e.g. an individual parameter β_j — and confidence interval

- Theorem 6: $\boldsymbol{p}^{\mathrm{T}}\boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta},\, \sigma^2 \boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}\right)$

  and $\dfrac{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{b} - \boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}} \sim t_{n-(k+1)}$ if $\mathrm{rank}(\boldsymbol{X}) = k + 1$

- *t*-test for an individual parameter β_j
- Confidence interval for the β_j

Multiple Linear Regression: Theorem 6

Theorem 6: Assume for simplicity that $\operatorname{rank}(\boldsymbol{X}) = k+1$

and let $\boldsymbol{p}^{\mathrm{T}} \in \mathbb{R}^{1\times(1+k)}$ be a non-zero row vector $(\boldsymbol{p}^{\mathrm{T}} \neq \boldsymbol{0}^{\mathrm{T}})$.

Then

$$\boldsymbol{p}^{\mathrm{T}}\boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}\right)$$

and

$$\frac{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{b} - \boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}} \sim t_{n-(k+1)}$$

Remark: The matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is positively definite.

Multiple Linear Regression: Prediction (Extrapolation)

Remark: Given a new row vector $\boldsymbol{x} = (x_0, x_1, x_2 \ldots, x_k) \in \mathbb{R}^{1\times(1+k)}$, which is not included in the matrix $\boldsymbol{X}$, we may wish to predict (extrapolate) the value of the random variable Y for this new statistical unit ($\boldsymbol{x}$).

Assuming that the model is true, we should have

$$Y_{\boldsymbol{x}} \approx \boldsymbol{x\beta}$$

or

$$Y_{\boldsymbol{x}} = \boldsymbol{x\beta} + \varepsilon$$

where ε is the random error.

Multiple Linear Regression: Prediction (Extrapolation)

Not knowing the parameters $\boldsymbol{\beta}$, we have to use their estimates $\boldsymbol{b}$ instead.

Then **the point estimate** of the value Y_x is:

$$\tilde{Y}_x = \boldsymbol{x}\boldsymbol{b}$$

Remark: If that $\mathrm{rank}(X) = k + 1$, then we can consider $\boldsymbol{p}^{\mathrm{T}} = \boldsymbol{x}$ and Theorem 6

$$\frac{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{b} - \boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}} \sim t_{n-(k+1)}$$

to obtain a confidence interval for the true value $\boldsymbol{x}\boldsymbol{\beta}$.

Multiple Linear Regression: Prediction (Extrapolation)

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $c > 0$ be the value such that

$$\int_{-c}^{+c} f(t)\,\mathrm{d}t = 1 - \alpha$$

where $f(t)$ is the density of Student's *t*-distribution with $n - (k+1)$ d.f.

Then, by Theorem 6,

$$P\left(-c \le \frac{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{b} - \boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}} \le +c\right) = P\left(-c \le \frac{\boldsymbol{x}\boldsymbol{b} - \boldsymbol{x}\boldsymbol{\beta}}{\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}} \le +c\right) = 1 - \alpha$$

Multiple Linear Regression: Prediction (Extrapolation)

By Theorem 6:

$$P\left(-c \le \frac{\boldsymbol{x}\boldsymbol{b} - \boldsymbol{x}\boldsymbol{\beta}}{\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}} \le +c\right) = 1 - \alpha$$

$$P\left(-c\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}} \le \boldsymbol{x}\boldsymbol{b} - \boldsymbol{x}\boldsymbol{\beta} \le +c\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}\right) = 1 - \alpha$$

$$P\left(\boldsymbol{x}\boldsymbol{b} - c\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}} \le \boldsymbol{x}\boldsymbol{\beta} \le \boldsymbol{x}\boldsymbol{b} + c\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}\right) = 1 - \alpha$$

Multiple Linear Regression: Prediction (Extrapolation)

SILESIAN UNIVERSITY

Having obtained

$$P\left(\boldsymbol{xb} - c\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}} \le \boldsymbol{x\beta} \le \boldsymbol{xb} + c\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}\right) = 1 - \alpha$$

the probability that the unknown

$$\boldsymbol{x\beta} \in \left[\boldsymbol{xb} - t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}},\ \boldsymbol{xb} + t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{\boldsymbol{p}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{p}}\right]$$

is about $1-\alpha = 95\ \%$,

where $t_{n-(k+1)}(q)$ denotes the **quantile function of Student's *t*-distribution**

Multiple Linear Regression: Theorem 6: Corollary

Corollary: By considering

$$p^{\mathrm{T}} = (0 \quad \ldots \quad 0 \quad 1 \quad 0 \quad \ldots \quad 0) \qquad \text{with the 1 at the } j\text{–th position}$$

we obtain:

$$\frac{b_j - \beta_j}{\sqrt{s^2}\sqrt{c_{jj}}} \sim t_{n-(k+1)} \qquad \text{for} \quad j = 0, 1, \ldots, k$$

Remark: Use the Corollary

— for *t*-tests about the parameters $\beta_0, \beta_1, \ldots, \beta_k$ of the model,

— to establish the confidence intervals for the parameters $\beta_0, \beta_1, \ldots, \beta_k$,

Tests of hypotheses about the individual parameters β_j

- Choose any non-zero $\boldsymbol{p}^{\mathrm{T}} \in \mathbb{R}^{1\times(1+k)}$ and let $a \in \mathbb{R}$ be a prescribed number.
- We can then use Theorem 5 to test the null hypothesis H_0 that $\boldsymbol{p}^{\mathrm{T}}\boldsymbol{\beta} = a$.

- By taking a particular choice of the non-zero $\boldsymbol{p}^{\mathrm{T}} \in \mathbb{R}^{1\times(1+k)}$, we can use

  the Corollary $\left(\frac{b_j-\beta_j}{\sqrt{s^2}\sqrt{c_{jj}}} \sim t_{n-(k+1)}\right)$ to test the null hypothesis H_0 that $\beta_j = a$

  or $\beta_j = 0$ (if we put $a = 0$ in particular).

t-test for the parameter β_j // rank(X)=k+1

Notation: Let

$$t_{n-(k+1)}(q)$$

denote the **quantile function of Student's *t*-distribution** with $n-(k+1)$ d.f.

The quantile function $t_{n-(k+1)}(q)$ is the function inverse to the cumulative distribution function $F(x)$ of **Student's *t*-distribution** with $n-(k+1)$ degrees of freedom, i.e.

$$t_{n-(k+1)}(q) = F^{-1}(q) \qquad \text{for} \quad q \in (0,1)$$

t-test for the parameter β_j // rank(X)=k+1

<u>Notation:</u> Let

$$t_{n-(k+1)}(q)$$

denote the **quantile function of Student's *t*-distribution** with $n-(k+1)$ d.f.

In other words, if $0 < q < 1$, then $x = t_{n-(k+1)}(q)$ is the unique value such that

$$\int_{-\infty}^{t_{n-(k+1)}(q)} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = q$$

where $f(t)$ is the density of Student's *t*-distribution with $n-(k+1)$ d.f.

t-test for the parameter β_j // rank(X)=k+1

Choosing the index $j \in \{0, 1, \dots, k\}$ and a value $b_{j0} \in \mathbb{R}$,

formulate the **null hypothesis**

$$H_0: \quad \beta_j = b_{j0}$$

formulate the **alternative hypothesis**

- two-sided: $H_1: \quad \beta_j \neq b_{j0}$
- one-sided: $H_1: \quad \beta_j < b_{j0}$
- one-sided: $H_1: \quad \beta_j > b_{j0}$

and use the aforementioned Corollary to conduct the test.

t-test for the parameter β_j // rank(X)=k+1

Having chosen the value $b_{j0} \in \mathbb{R}$, such as $b_{j0} = 0$, and assuming the null hypothesis H_0: $\beta_j = b_{j0}$ is true, calculate the statistic

$$T = \frac{b_j - \beta_j}{\sqrt{s^2}\sqrt{c_{jj}}} = \frac{b_j - b_{j0}}{\sqrt{s^2}\sqrt{c_{jj}}} = \frac{b_j}{\sqrt{s^2}\sqrt{c_{jj}}}$$

t-test for the parameter β_j // rank(X)=k+1

The t-test for β_j with two-sided alternative hypothesis ($\beta_j \neq b_{j0}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\ \%$, other popular values are $\alpha = 10\ \%$ or $\alpha = 1\ \%$ or $\alpha = 0.1\ \%$ etc.
- the **critical value** is $c = t_{n-(k+1)}\left(1 - \frac{\alpha}{2}\right)$
- if $T \in (-\infty, -c] \cup [+c, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-c, +c)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β_j // rank(X)=k+1

The t-test for β_j with one-sided alternative hypothesis ($\beta_j < b_{j0}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.

- the **critical value** is $c = t_{n-(k+1)}(1-\alpha)$

- if $T \in (-\infty, -c]$, **the critical region**, then **reject** the null hypothesis

- if $T \in (-c, +\infty)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β_j // rank(X)=k+1

The t-test for β_j with one-sided alternative hypothesis ($\beta_j > b_{j0}$):

- choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$, other popular values are $\alpha = 10\,\%$ or $\alpha = 1\,\%$ or $\alpha = 0.1\,\%$ etc.
- the **critical value** is $c = t_{n-(k+1)}(1-\alpha)$
- if $T \in [+c, +\infty)$, **the critical region**, then **reject** the null hypothesis
- if $T \in (-\infty, +c)$, then **do not reject** (or fail to reject) the null hypothesis

t-test for the parameter β_j // rank(X)=k+1

¡¡¡WARNING!!! It usually makes no sense to test the null hypothesis $\beta_0 = 0$ if $X_0 = 1$, that is, if β_0 is the intercept term. Do not use the aforementioned test unless you know what and why you are doing.

It can make sense to test the null hypothesis $\beta_0 = 0$ if the independent values $x_1, x_2, \ldots, x_n$ are from a neighbourhood of zero.

Otherwise (if the cluster of $x_1, x_2, \ldots, x_n$ is far from zero) it hardly makes sense to test the null hypothesis $\beta_0 = 0$ because the intercept term β_0 is just a constant

Confidence interval for the parameter β_j // rank(X)=k+1

Let $x, y \in \mathbb{R}$ be any numbers such that $x < y$ and let $F(x)$ be the cumulative distribution function of Student's *t*-distribution with $n-(k+1)$ degrees of freedom. Then, by the definition of the cumulative distribution function and by the Corollary, the probability

$$P\left(x < \frac{b_j - \beta_j}{\sqrt{s^2}\sqrt{c_{jj}}} \leq y\right) = F(y) - F(x)$$

Therefore

$$P\left(x\sqrt{s^2}\sqrt{c_{jj}} < b_j - \beta_j \leq y\sqrt{s^2}\sqrt{c_{jj}}\right) = F(y) - F(x)$$

$$P\left(b_j - y\sqrt{s^2}\sqrt{c_{jj}} \leq \quad \beta_j \quad < b_j - x\sqrt{s^2}\sqrt{c_{jj}}\right) = F(y) - F(x)$$

Confidence interval for the parameter β_j // rank(X)=k+1

We have:

$$P\left(b_j - y\sqrt{s^2}\sqrt{c_{jj}} \le \beta_j < b_j - x\sqrt{s^2}\sqrt{c_{jj}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\ \%$.

Let $y = t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right)$ and let $x = -y = -t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right) = t_{n-(k+1)}\left(\frac{\alpha}{2}\right)$.

Recall that $t_{n-(k+1)}(q) = F^{-1}(q)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta_j \in \left[b_j - t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{c_{jj}},\ b_j + t_{n-(k+1)}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2}\sqrt{c_{jj}}\right]$$

Confidence interval for the parameter β_j // rank(X)=k+1

We have:

$$P\left(b_j - y\sqrt{s^2}\sqrt{c_{jj}} \leq \beta_j < b_j - x\sqrt{s^2}\sqrt{c_{jj}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = t_{n-(k+1)}(1-\alpha)$ and let $x = -\infty$. Recall that $t_{n-(k+1)}(q) = F^{-1}(q)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

$$\text{the unknown } \beta_j \in \left[b_j - t_{n-(k+1)}(1-\alpha)\sqrt{s^2}\sqrt{c_{jj}},\ +\infty\right)$$

is about $1-\alpha = 95\ \%$.

Confidence interval for the parameter β_j // rank(X)=k+1

We have:

$$P\left(b_j - y\sqrt{s^2}\sqrt{c_{jj}} \le \beta_j < b_j - x\sqrt{s^2}\sqrt{c_{jj}}\right) = F(y) - F(x)$$

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Let $y = +\infty$ and let $x = t_{n-(k+1)}(\alpha)$. Recall that $t_{n-(k+1)}(q) = F^{-1}(q)$.

Then, by the continuity of the cumulative distribution function, **the probability** that

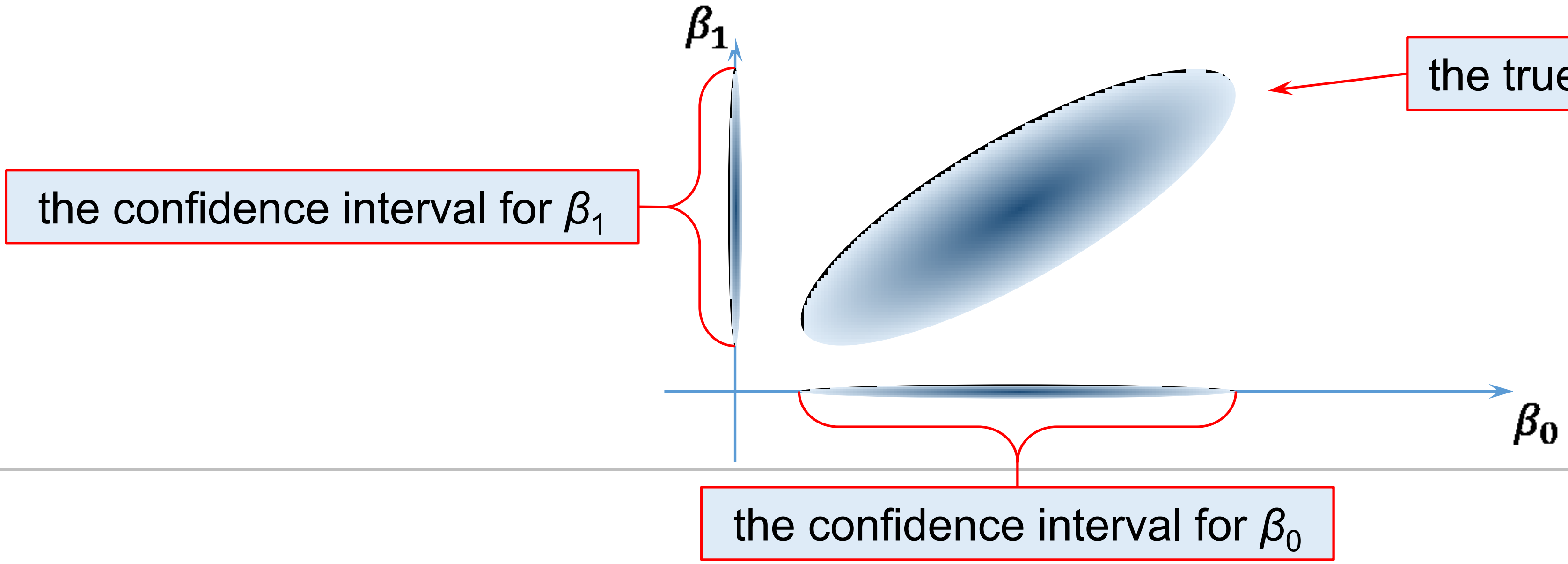
$$\text{the unknown } \beta_j \in \left(-\infty,\ b_j + t_{n-(k+1)}(\alpha)\sqrt{s^2}\sqrt{c_{jj}}\right]$$

is about $1-\alpha = 95\,\%$.

Confidence interval for the parameter β_j // rank(X)=k+1

¡¡¡ WARNING !!!

- **Never use the above *t*-test for the parameters $\beta_0,\beta_1,\ldots,\beta_k$ consecutively!**
- **Never use the above construction of the confidence intervals consecutively!**
- **Use the following result (Theorem 7) instead !**



F-test for the significance of the model and confidence region & *F*-test for a system of linear combinations of the parameters $\beta_0,\beta_1,\ldots,\beta_k$

- Theorem 7:

$$\frac{(\boldsymbol{b}-\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X})(\boldsymbol{b}-\boldsymbol{\beta})}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} \sim F_{k+1,\,n-(k+1)}$$

- *F*-test for the significance of the model
- Confidence region
- Theorem 8:

$$\frac{(\boldsymbol{A}\boldsymbol{b}-\boldsymbol{a})^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^{\mathrm{T}}\right)^{-1}(\boldsymbol{A}\boldsymbol{b}-\boldsymbol{a})}{\mathrm{RSS}} \Big/ \frac{r}{n-(k+1)} \sim F_{r,\,n-(k+1)}$$

Multiple Linear Regression: Theorem 7

Theorem 7: Assume for simplicity that $\operatorname{rank}(X) = k + 1$.

It holds

$$\frac{(\boldsymbol{b}-\boldsymbol{\beta})^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)(\boldsymbol{b}-\boldsymbol{\beta})}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} \sim F_{k+1,\, n-(k+1)}$$

Multiple Linear Regression: Theorem 7*

<u>Theorem 7*:</u> Assume for simplicity that $\operatorname{rank}(\boldsymbol{X}) = k + 1$.

Let $\boldsymbol{a} \in \mathbb{R}^{1+k}$ be a vector.

If

$$\boldsymbol{\beta} = \boldsymbol{a}$$

then

$$\frac{(\boldsymbol{b}-\boldsymbol{a})^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)(\boldsymbol{b}-\boldsymbol{a})}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} \sim F_{k+1,\,n-(k+1)}$$

Multiple Linear Regression: Theorem 7*: Corollary

Corollary: By considering

$$\boldsymbol{a} = \boldsymbol{0}$$

that is the zero vector, we are testing the null hypothesis that

$$H_0: \quad \boldsymbol{\beta} = \boldsymbol{0}$$

that is

$$H_0: \quad \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

that is **we are testing the overall significance of the model**.

Multiple Linear Regression: Theorem 7*: Corollary

Corollary: By considering

$$\boldsymbol{a} = \boldsymbol{0}$$

we obtain:

$$\frac{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b}}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} = \frac{\hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}}}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} \sim F_{k+1,\, n-(k+1)}$$

Remark: Use this Corollary

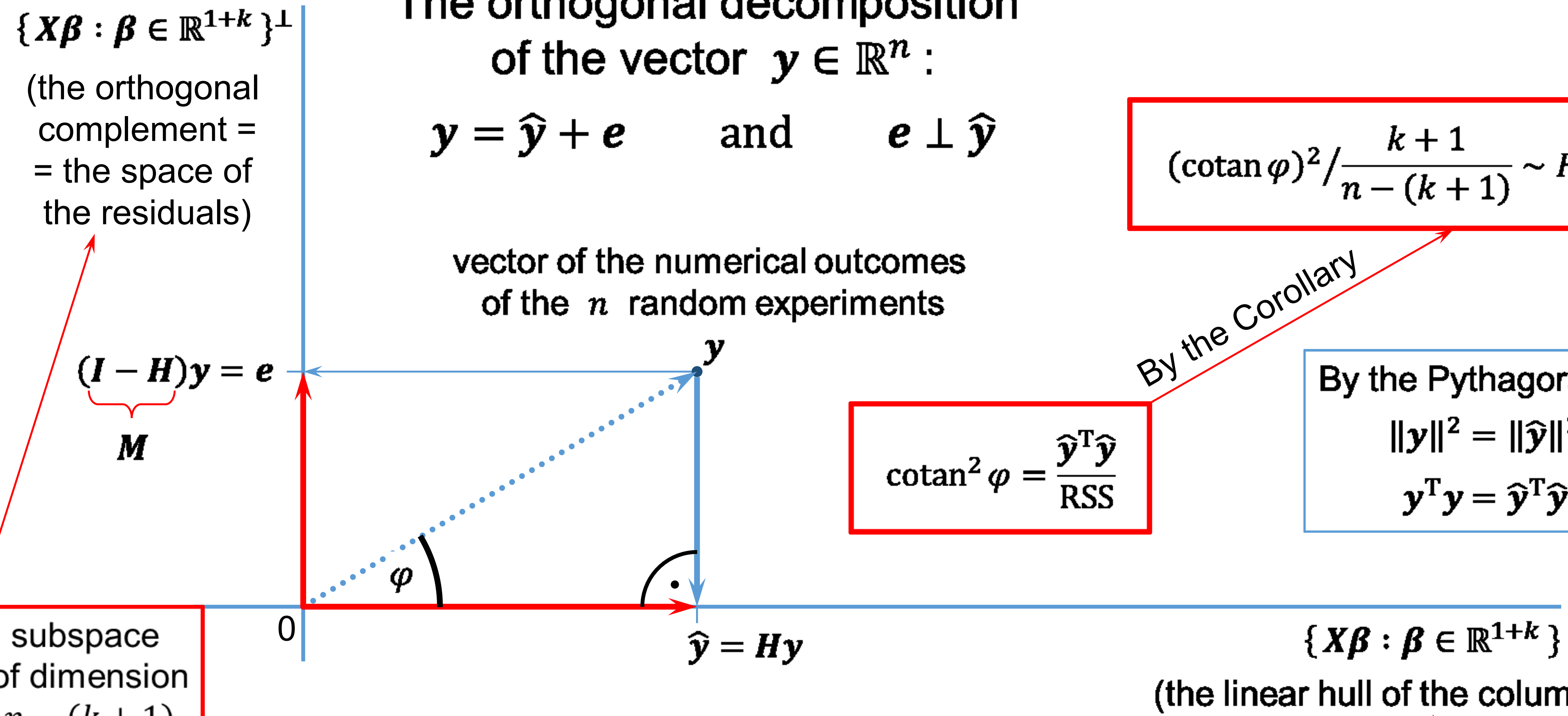
— for *F*-test about the significance of the model,

— to establish the confidence region.

Multiple Linear Regression: Theorem 7*: Corollary

The orthogonal decomposition of the vector $\boldsymbol{y} \in \mathbb{R}^n$:

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} \qquad \text{and} \qquad \boldsymbol{e} \perp \hat{\boldsymbol{y}}$$



subspace of dimension $n - (k+1)$

subspace of dimension $k + 1$

$$\text{cotan}^2\,\varphi = \frac{\hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}}}{\text{RSS}}$$

By the Corollary

$$(\text{cotan}\,\varphi)^2 \Big/ \frac{k+1}{n-(k+1)} \sim F_{k+1,\,n-(k+1)}$$

By the Pythagoras Theorem:

$$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{y}}\|^2 + \|\boldsymbol{e}\|^2$$

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = \hat{\boldsymbol{y}}^{\mathrm{T}}\hat{\boldsymbol{y}} + \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$

Residual Sum of Squares: $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$

F-test for the significance of the model // rank(*X*)=*k*+1

<u>Notation:</u> Let

$$F_{k+1,n-(k+1)}(q)$$

denote the **quantile function of Fisher's *F*-distribution** with $k+1$ and $n-(k+1)$ degrees of freedom.

The quantile function $F_{k+1,n-(k+1)}(q)$ is the function inverse to the cumulative distribution function $F(x)$ of **Fisher's *F*-distribution** with $k+1$ and $n-(k+1)$ degrees of freedom, i.e.

$$F_{k+1,n-(k+1)}(q) = F^{-1}(q) \qquad \text{for} \quad q \in (0,1)$$

F-test for the significance of the model // rank(X)=k+1

<u>Notation:</u> Let

$$F_{k+1,n-(k+1)}(q)$$

denote the **quantile function of Fisher's *F*-distribution** with $k+1$ and $n-(k+1)$ degrees of freedom.

In other words, if $0 < q < 1$, then $x = F_{k+1,n-(k+1)}(q)$ is the unique value such that

$$\int_{-\infty}^{F_{k+1,n-(k+1)}(q)} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = q$$

where $f(t)$ is the density of Fisher's *F*-distribution with $k+1$ and $n-(k+1)$ d.f.

F-test for the significance of the model // rank(X)=k+1

Formulate the **null hypothesis**

$$H_0: \qquad \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

- **Be cautious because it usually makes no sense to test the value of the intercept term** β_0 (see above).
  See also the Coefficient of Determination (R^2) below.

- The alternative hypothesis is simply $H_1 \equiv \neg H_0$, the logical negation of H_0, that is $\beta_j \neq 0$ for at least one $j \in \{0, 1, \dots, k\}$.

F-test for the significance of the model // rank(*X*)=*k*+1

- Calculate the statistic

$$F = \frac{\hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}}}{\mathrm{RSS}} \bigg/ \frac{k+1}{n-(k+1)}$$

- Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

- The **critical value** is $c = F_{k+1,n-(k+1)}(1-\alpha)$, that is $\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$, where $f(x)$ is the density of Fisher's *F*-distribution with $k+1$ and $n-(k+1)$ degrees of freedom,

- If $F \in [c, +\infty)$, the **critical region**, then **reject** the null hypothesis.

- If $F \in (0, c)$, then **do not reject** (fail to reject) the null hypothesis.

Confidence region for the parameters // rank(X)=k+1

Consider $\overline{\boldsymbol{\beta}} = \mathbf{0}$, let $x \in \mathbb{R}$ be any real number, and let $F(x)$ be the cumulative distribution function of Fisher's *F*-distribution with $k+1$ and $n-(k+1)$ degrees of freedom. Then, by the definition of the cumulative distribution function and by the Corollary, the probability

$$P\left(\frac{(\boldsymbol{b}-\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{b}-\boldsymbol{\beta})}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} \leq x\right) = F(x) \qquad \text{for any} \quad \boldsymbol{\beta} \in \mathbb{R}^{1+k}$$

Confidence region for the parameters // rank(X)=k+1

Choose the level of significance, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.

Then the probability that

the unknown $$\boldsymbol{\beta} \in \left\{ \boldsymbol{\beta} \in \mathbb{R}^{1+k} : \frac{(\boldsymbol{b}-\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} (\boldsymbol{b}-\boldsymbol{\beta})}{\mathrm{RSS}} \Big/ \frac{k+1}{n-(k+1)} \leq F_{k+1,\, n-(k+1)}(1-\alpha) \right\}$$

is about $1-\alpha = 95\,\%$.

Remark: This confidence region is an ellipsoid centred at $\boldsymbol{b}$.

The nominator $((\boldsymbol{b}-\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} (\boldsymbol{b}-\boldsymbol{\beta}))$ is a quadratic expression in $\boldsymbol{\beta}$.

To gain a geometrical insight, calculate the spectral / eigendecomposition

Multiple Linear Regression: Theorem 8

Theorem 8: Assume for simplicity that $\operatorname{rank}(\boldsymbol{X}) = k + 1$. Let $\boldsymbol{a} \in \mathbb{R}^r$ be a vector and let $\boldsymbol{A} \in \mathbb{R}^{r \times (1+k)}$ be an $r \times (1 + k)$ matrix of full-rank where $r \leq 1 + k$, that is

$$r = \operatorname{rank}(\boldsymbol{A}) \leq \operatorname{rank}(\boldsymbol{X}) = k + 1$$

If

$$\boldsymbol{A\beta} = \boldsymbol{a}$$

then

$$\frac{(\boldsymbol{Ab} - \boldsymbol{a})^{\mathrm{T}}\left(\boldsymbol{ACA}^{\mathrm{T}}\right)^{-1}(\boldsymbol{Ab} - \boldsymbol{a})}{\mathrm{RSS}} \Big/ \frac{r}{n - (k + 1)} \sim F_{r,n-(k+1)}$$

Multiple Linear Regression: Theorem 8: Illustration

It holds:

$$(\boldsymbol{Ab}-\boldsymbol{a})^{\mathrm{T}}\left(\boldsymbol{ACA}^{\mathrm{T}}\right)^{-1}(\boldsymbol{Ab}-\boldsymbol{a}) = \|\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}'\|^2 =$$

$$= (\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}')^{\mathrm{T}}(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}$$

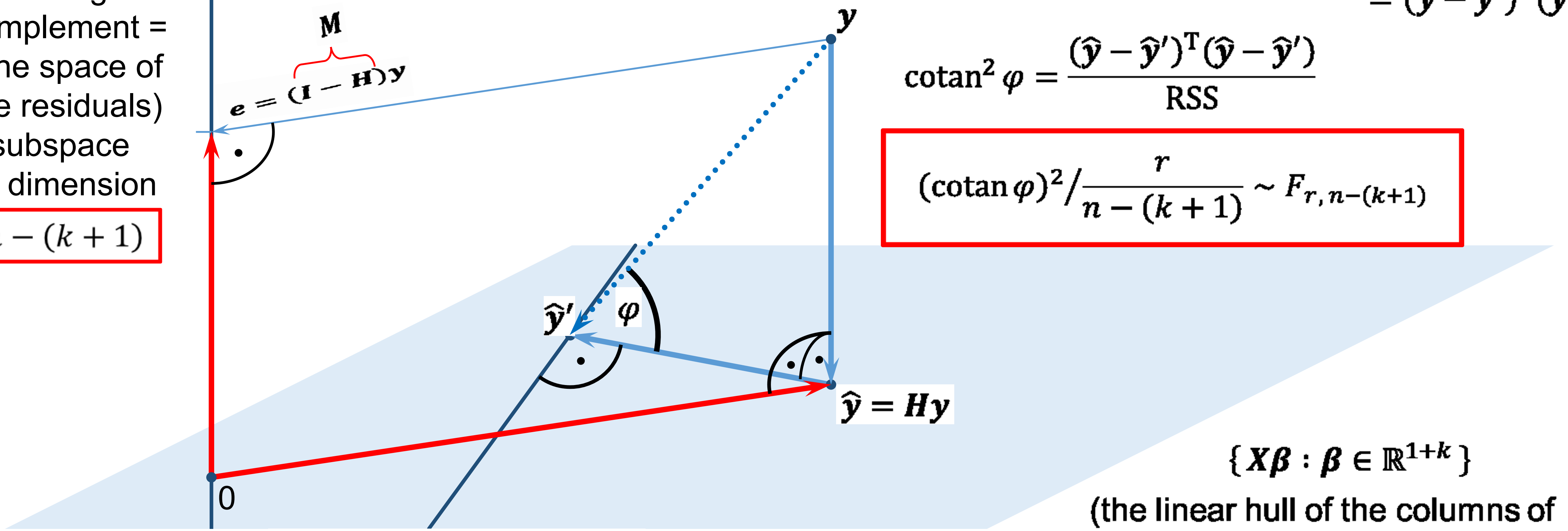
$$\text{cotan}^2\,\varphi = \frac{(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}')^{\mathrm{T}}(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}')}{\text{RSS}}$$

$$(\text{cotan}\,\varphi)^2 \Big/ \frac{r}{n-(k+1)} \sim F_{r,\,n-(k+1)}$$

$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}^{\perp}$

(the orthogonal complement = = the space of the residuals) subspace of dimension

$n-(k+1)$



$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$

(the linear hull of the columns of $\boldsymbol{X}$) subspace of dimension

$k+1$

this is an affine subspace of dimension $k+1-r$ → $\{\boldsymbol{X\beta} : \boldsymbol{A\beta} = \boldsymbol{a},\ \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$

the dimension of its complement within the subspace of dimension $k+1$ is r

Multiple Linear Regression: Theorem 8: Remark

Theorem 8

$$A\beta = a \quad \Rightarrow \quad \frac{(Ab - a)^{\mathrm{T}}\left(ACA^{\mathrm{T}}\right)^{-1}(Ab - a)}{\mathrm{RSS}} \bigg/ \frac{r}{n - (k+1)} \sim F_{r, n-(k+1)}$$

is at the heart of the ANOVA method

and other results.

The Coefficient of Determination (R^2)

- Assumption: $\mathbf{1} \in \{ \boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k} \}$
- Motivation
- Some facts
- Theorem 8: Corollary
- *F*-test for the null hypothesis

  H_0: $\beta_1 = \ldots = \beta_k = 0$

The Coefficient of Determination (R^2): Assumption

Assume throughout this section that

$$\mathbf{1} \in \{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$$

where $\mathbf{1}$ is the vector of n ones:

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

For example, assume that

$$X_0 = \mathbf{1}$$

that is β_0 is the intercept term.

The Coefficient of Determination (R^2): Th. 8: Corollary

If $X_0 = 1$, that is β_0 is the intercept term, then it may be desirable to test the null hypothesis

$$H_0: \quad \beta_1 = \cdots = \beta_k = 0$$

that is without the test for the parameter β_0.

To this end, apply Theorem 8 with the $k \times (k+1)$ matrix and the k-vector

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{a} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The Coefficient of Determination (R^2): Th. 8: Corollary

Recall our assumption that

$$\mathbf{1} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$$

Then the line

$$\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\} \subset \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$$

In particular, if $X_0 = 1$, that is

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ 1 & x_{31} & \dots & x_{3k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \qquad \boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{a} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

then

The Coefficient of Determination (R^2): Th. 8: Corollary

Let: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

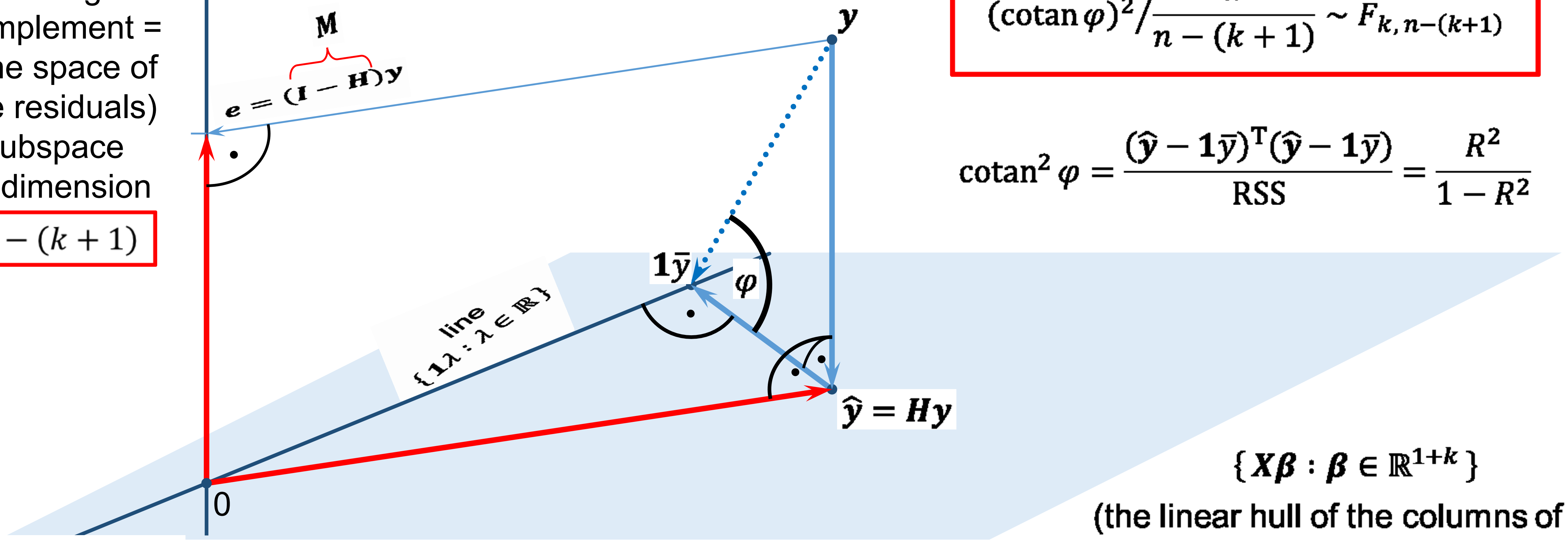
By Theorem 8:

$$(\operatorname{cotan}\varphi)^2 / \frac{k}{n-(k+1)} \sim F_{k,\,n-(k+1)}$$

$$\operatorname{cotan}^2\varphi = \frac{(\hat{\mathbf{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\mathbf{y}} - \mathbf{1}\bar{y})}{\mathrm{RSS}} = \frac{R^2}{1-R^2}$$

$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}^{\perp}$

(the orthogonal complement = = the space of the residuals) subspace of dimension $n-(k+1)$



$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$

(the linear hull of the columns of $\boldsymbol{X}$) subspace of dimension $k+1$

$\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$

the line is a subspace of dimension 1

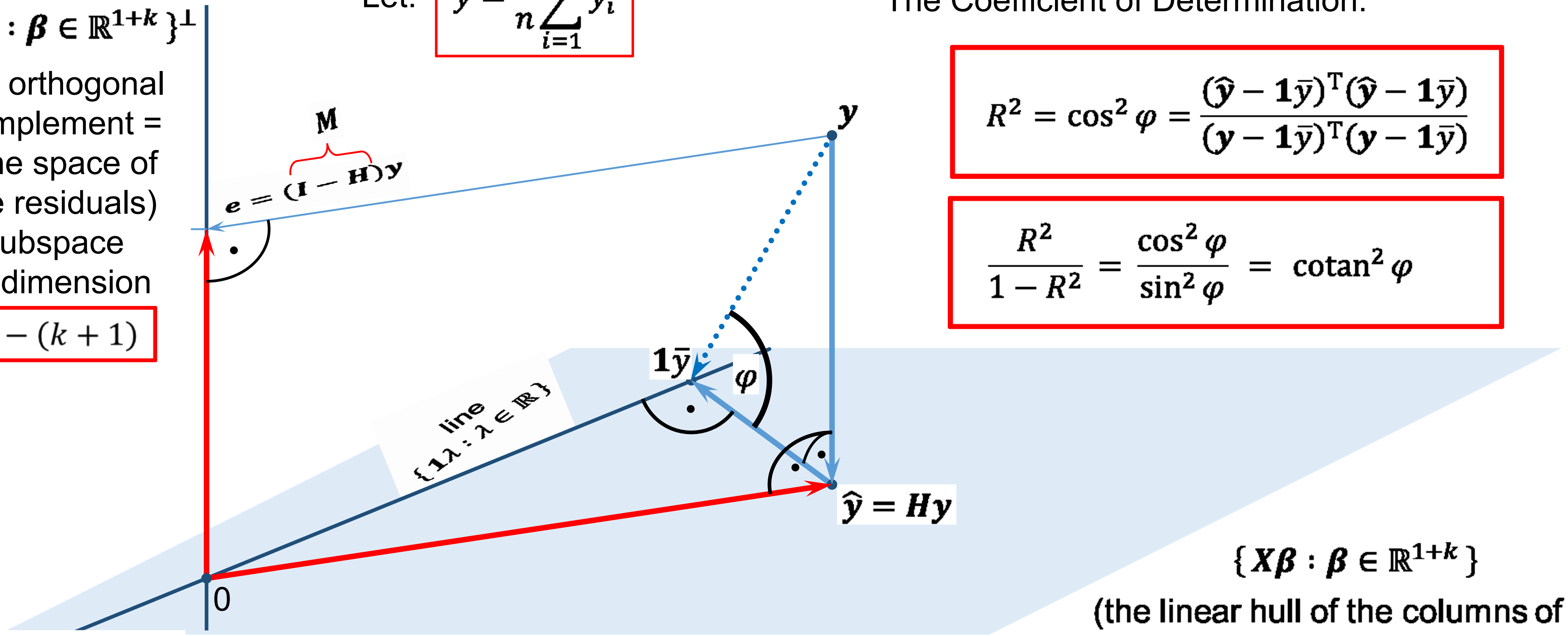
the dimension of its complement within the subspace of dimension $k+1$ is k

The Coefficient of Determination (R^2): Th. 8: Corollary

Let: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}^{\perp}$

(the orthogonal complement = = the space of the residuals) subspace of dimension $n - (k+1)$



The Coefficient of Determination:

$$R^2 = \cos^2\varphi = \frac{(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})}{(\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y})}$$

$$\frac{R^2}{1 - R^2} = \frac{\cos^2\varphi}{\sin^2\varphi} = \operatorname{cotan}^2\varphi$$

$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$

(the linear hull of the columns of $\boldsymbol{X}$) subspace of dimension $k + 1$

$\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$

the line is a subspace of dimension 1

the dimension of its complement within the subspace of dimension $k+1$ is k

The Coefficient of Determination (R^2): TSS=RSS+RegSS

Introduce the **Total Sum of Squares:**

$$\mathrm{TSS} = (\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y}) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Introduce the **Regression Sum of Squares:**

$$\mathrm{RegSS} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Recall the **Residual Sum of Squares:**

$$\mathrm{RSS} = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \sum_{i=1}^{n} e_i^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The Coefficient of Determination (R^2): TSS=RSS+RegSS

Let: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

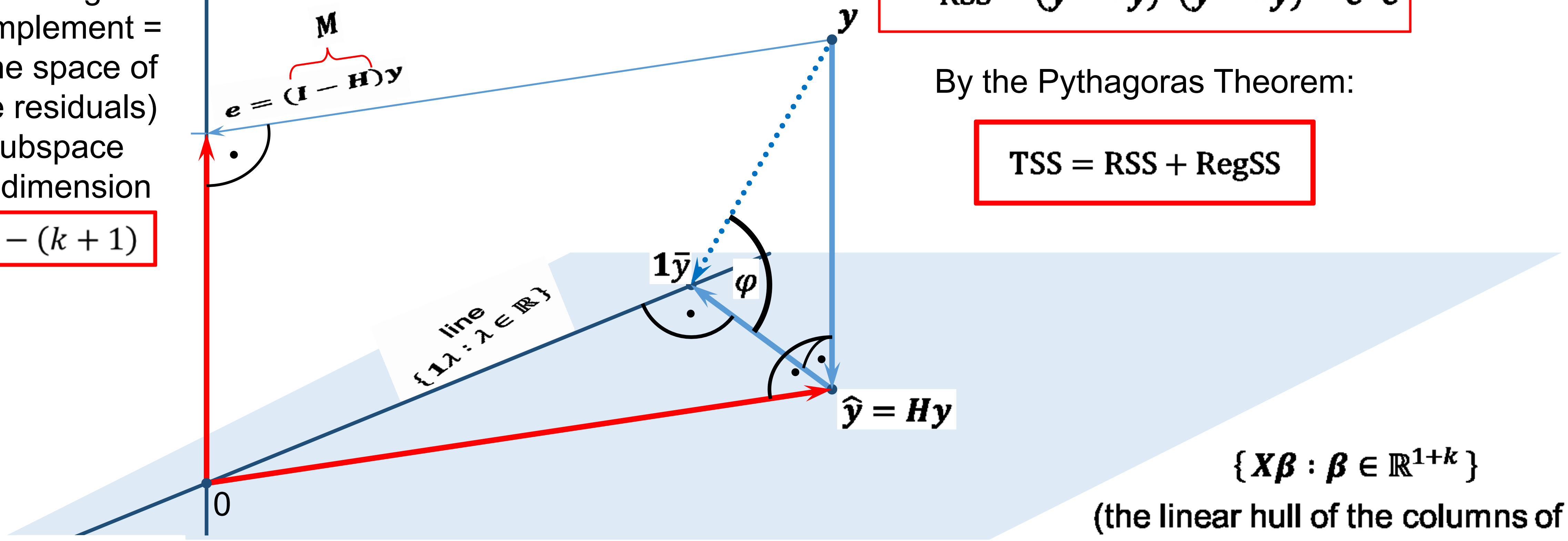
$$\mathrm{TSS} = (\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y})$$

$$\mathrm{RegSS} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$$

$$\mathrm{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$

$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}^{\perp}$

(the orthogonal complement = = the space of the residuals) subspace of dimension $n-(k+1)$



By the Pythagoras Theorem:

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{RegSS}$$

$\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$

(the linear hull of the columns of $\boldsymbol{X}$) subspace of dimension $k+1$

$\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$

the line is a subspace of dimension 1

the dimension of its complement within the subspace of dimension $k+1$ is k

The Coefficient of Determination (R^2): Some facts

Proposition: Under the assumption $\mathbf{1} \in \{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$, it holds

$$\mathbf{1}^{\mathrm{T}}\mathbf{1}\bar{y} = \mathbf{1}^{\mathrm{T}}\boldsymbol{y} = \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{y}}$$

In words:

All three points $\mathbf{1}\bar{y}$, $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$ lie in the hyperplane

$$\{\boldsymbol{\beta} \in \mathbb{R}^{1+k} : \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{1}^{\mathrm{T}}\mathbf{1}\bar{y}\}$$

which is perpendicular to the line $\{\mathbf{1}\lambda : \lambda \in \mathbb{R}\}$.

The Coefficient of Determination (R^2): Some facts

Proposition: Under the assumption $\mathbf{1} \in \{ \boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k} \}$, it holds

$$\mathbf{1}^{\mathrm{T}}\mathbf{1}\bar{y} = \mathbf{1}^{\mathrm{T}}\boldsymbol{y} = \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{y}}$$

Corollary:

$$\mathbf{1}^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \mathbf{1}^{\mathrm{T}}\boldsymbol{e} = \sum_{i=1}^{n} e_i = 0$$

The Coefficient of Determination (R^2): Some facts

Proposition: Under the assumption $\mathbf{1} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$, it holds

$$\mathbf{1}^{\mathrm{T}}\mathbf{1}\bar{y} = \mathbf{1}^{\mathrm{T}}\boldsymbol{y} = \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{y}}$$

Proof:

The assumption equivalently says $\boldsymbol{H}\mathbf{1} = \mathbf{1}$, where $\boldsymbol{H}$ is the matrix

of the orthogonal projection onto the subspace $\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$.

Recall the matrix $\boldsymbol{H}$ is symmetric ($\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{H}$), therefore $\mathbf{1}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}} = \mathbf{1}^{\mathrm{T}}\boldsymbol{H} = \mathbf{1}^{\mathrm{T}}$.

Therefore $\mathbf{1}^{\mathrm{T}}\boldsymbol{y} = \mathbf{1}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{y} = \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{y}}$.

The first equality is obvious:

$\mathbf{1}^{\mathrm{T}}\mathbf{1}\bar{y} = \sum_{i=1}^{m} 1 \times 1 \times \sum_{i=1}^{n} y_i/n = n \times \sum_{i=1}^{n} y_i/n = \sum_{i=1}^{n} y_i = \mathbf{1}^{\mathrm{T}}\boldsymbol{y}$.

The Coefficient of Determination (R^2): TSS=RSS+RegSS

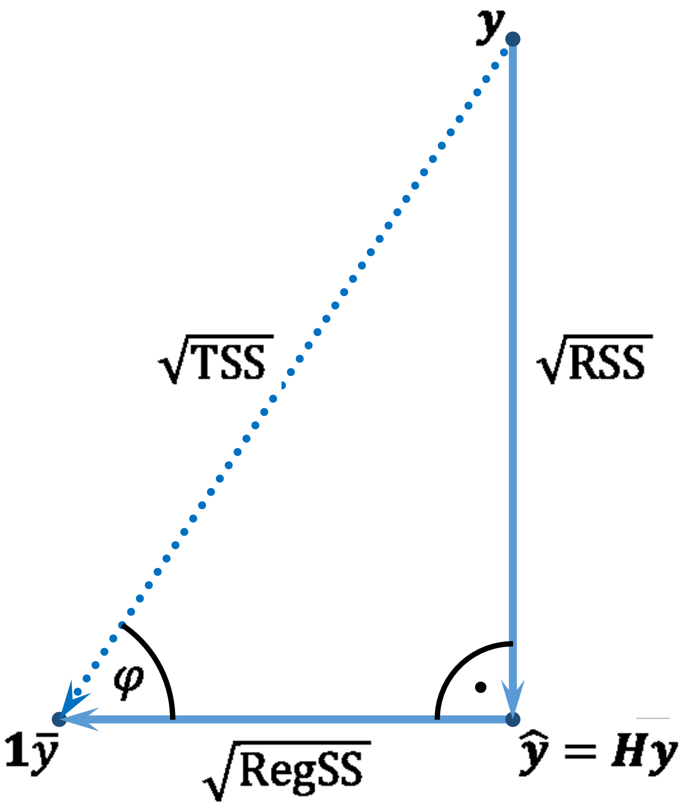
Proposition: Under the assumption $\mathbf{1} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$, it holds

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{RegSS}$$

$$\mathrm{TSS} = (\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y})$$
$$\mathrm{RegSS} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$$
$$\mathrm{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$

Proof:

The point $\hat{\boldsymbol{y}}$ is the orthogonal projection of the point $\boldsymbol{y}$ onto $\{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$, therefore $(\boldsymbol{y} - \hat{\boldsymbol{y}}) \perp \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$. We have $\hat{\boldsymbol{y}} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$, and we assume $\mathbf{1} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$, whence $\mathbf{1}\bar{y} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$ follows, therefore $\hat{\boldsymbol{y}} - \mathbf{1}\bar{y} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$ and $(\boldsymbol{y} - \hat{\boldsymbol{y}}) \perp (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$. By using the Pythagoras Theorem,



The Coefficient of Determination (R^2): Some facts

Proposition: Under the assumption $\mathbf{1} \in \{\boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$,

$$\mathrm{TSS} = (\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y})$$

$$\mathrm{RegSS} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$$

$$\mathrm{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$

it holds

$$(\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$$

Proof:

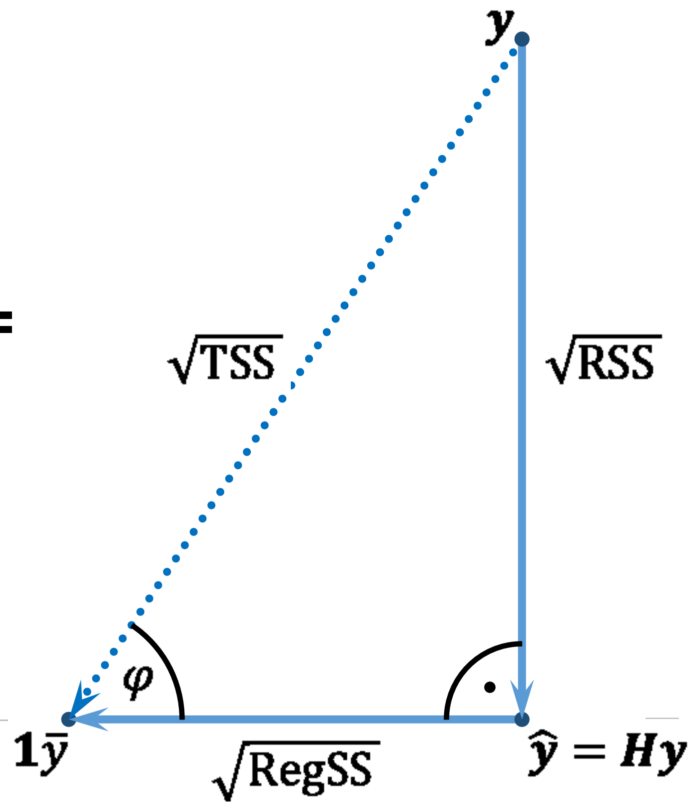
$$(\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) = \left((\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})\right)^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) =$$

$$= (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) + (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) =$$

$$= 0 + (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) =$$

$$= (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$$

q.e.d.



The Coefficient of Determination (R^2)

Assuming $1 \in \{X\beta : \beta \in \mathbb{R}^{1+k}\}$, define the

Coefficient of Determination:

$$R^2 = \frac{[(y - 1\bar{y})^T(\hat{y} - 1\bar{y})]^2}{(y - 1\bar{y})^T(y - 1\bar{y}) \times (\hat{y} - 1\bar{y})^T(\hat{y} - 1\bar{y})} =$$

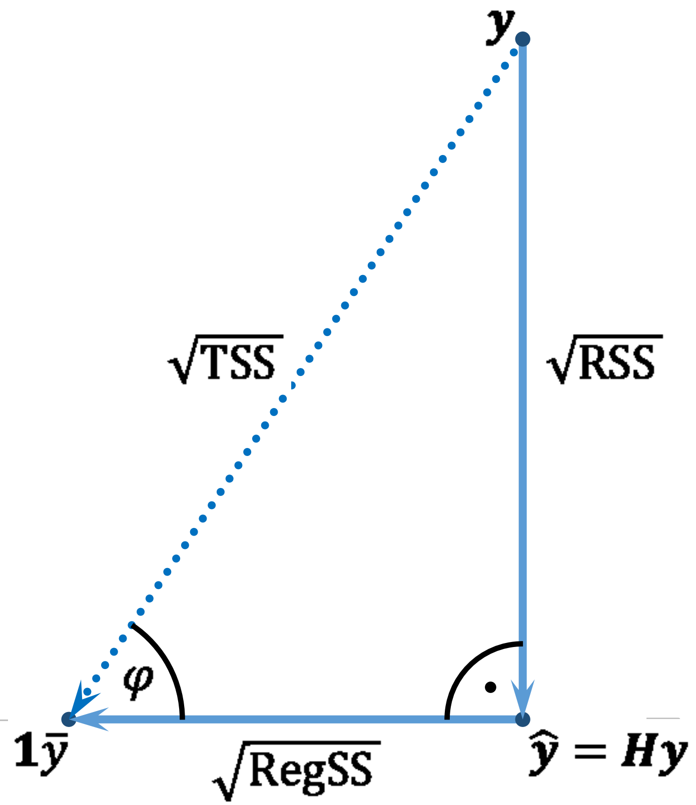
$$= \frac{[(\hat{y} - 1\bar{y})^T(\hat{y} - 1\bar{y})]^2}{(y - 1\bar{y})^T(y - 1\bar{y}) \times (\hat{y} - 1\bar{y})^T(\hat{y} - 1\bar{y})} =$$

$$= \frac{(\hat{y} - 1\bar{y})^T(\hat{y} - 1\bar{y})}{(y - 1\bar{y})^T(y - 1\bar{y})} = \cos^2\varphi = \frac{\text{RegSS}}{\text{TSS}}$$

$$\text{TSS} = (y - 1\bar{y})^T(y - 1\bar{y})$$

$$\text{RegSS} = (\hat{y} - 1\bar{y})^T(\hat{y} - 1\bar{y})$$

$$\text{RSS} = (y - \hat{y})^T(y - \hat{y}) = e^T e$$



The Coefficient of Determination (R^2)

Assuming $\mathbf{1} \in \{ \boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k} \}$, define the

Coefficient of Determination:

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

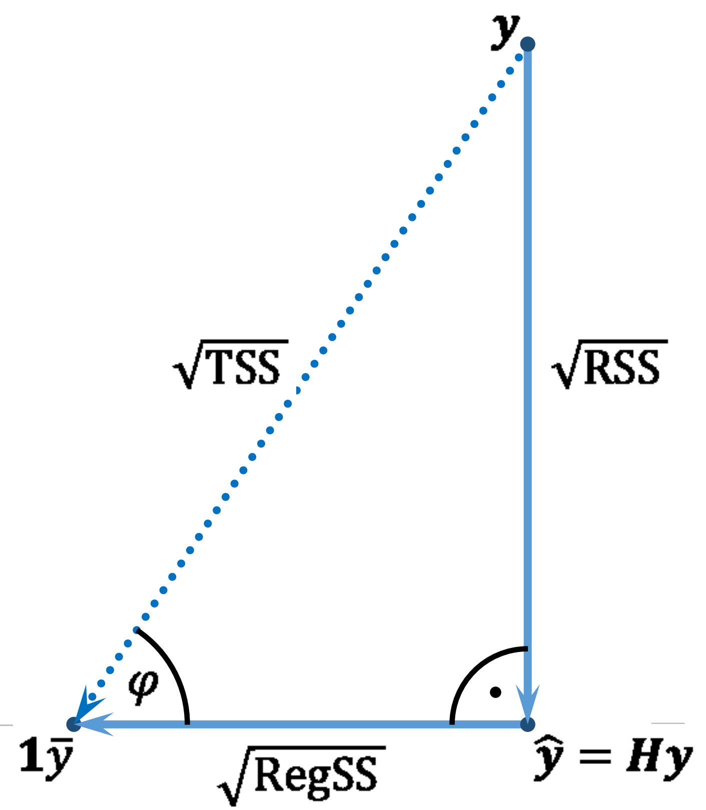
$$R^2 = \cos^2 \varphi = \frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{cotan}^2 \varphi = \frac{\cos^2 \varphi}{\sin^2 \varphi} = \frac{R^2}{1 - R^2} = \frac{\text{RegSS}}{\text{RSS}}$$

$$\text{TSS} = (\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y})$$

$$\text{RegSS} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})$$

$$\text{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$



The Coefficient of Determination (R^2): Th. 8: Corollary

Theorem 8: Corollary: Assume for simplicity that $\operatorname{rank}(X) = k + 1$ and assume that $1 \in \{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^{1+k}\}$. Under the hypothesis that

$$\beta_1 = \cdots = \beta_k = 0$$

it holds

$$(\operatorname{cotan} \varphi)^2 \Big/ \frac{k}{n-(k+1)} = \frac{R^2}{1-R^2} \Big/ \frac{k}{n-(k+1)} \sim F_{k,n-(k+1)}$$

$$= \frac{\text{RegSS}}{\text{RSS}} \Big/ \frac{k}{n-(k+1)} \sim F_{k,n-(k+1)}$$

F-test for the null hypothesis H_0: $\beta_1 = \ldots = \beta_k = 0$

- Choose **the level of significance**, a small number $\alpha > 0$, such as $\alpha = 5\,\%$.
- Find the **critical value** $c > 0$ so that $\int_c^{+\infty} f(x)\,\mathrm{d}x = \alpha$, where f is the density of the *F*-distribution with k and $n-(k+1)$ degrees of freedom.
- Calculate the statistic

$$F = \frac{R^2}{1-R^2} \Big/ \frac{k}{n-(k+1)} = \frac{\mathrm{RegSS}}{\mathrm{RSS}} \Big/ \frac{k}{n-(k+1)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \Big/ \frac{k}{n-(k+1)}$$

- If $F \in [c, +\infty)$, **the critical region**, then **reject** the null hypothesis.
- If $F \in [0, c)$, then **do not reject** (or fail to reject) the null hypothesis.

The Coefficient of Determination (R^2)

Remark: The above *F*-test is one-factor ANOVA in fact.

The coefficient of determination

$$R^2 = \cos^2\varphi = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{RegSS}}{\mathrm{TSS}}$$

is a "measure" (?) "how well the regression hyperplane $Y = b_0 + b_1X_1 + \cdots + b_kX_k$ fits the observed data (y_1, x_1), (y_2, x_2), …, (y_n, x_n)".

It holds

$$0 \le R^2 \le 1$$

The Coefficient of Determination (R^2)

$$R^2 = \cos^2 \varphi$$

If $R^2 \nearrow 1$

- then

$$F = \frac{R^2}{1 - R^2} \Big/ \frac{k}{n - (k+1)} \nearrow +\infty$$

- then
  - reject the null hypothesis that $(\beta_1 = \cdots = \beta_k = 0)$

- then
  - say "the fit is good"

The Coefficient of Determination (R^2)

$$R^2 = \cos^2 \varphi$$

If $R^2 \searrow 0$

- then

$$F = \frac{R^2}{1 - R^2} \Big/ \frac{k}{n - (k + 1)} \searrow 0$$

- then
  - fail to reject the null hypothesis that $(\beta_1 = \cdots = \beta_k = 0)$
  - it may be the case that

$$\mathrm{E}[y_i] = \beta_0 \qquad \text{for all} \quad i = 1, 2, \dots, n$$ (cf. ANOVA)

  - the sample (y_1, x_1), (y_2, x_2) …, (y_n, x_n) may come from one population
- then