Statistical Methods for Economists

Lecture (7 & 8)a

One-Way Analysis of Variance
(ANOVA)

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistical Methods for Economists
INM/BASTE

Outline of the lecture

- One-Way ANOVA: Introduction and Motivation
- One-Way ANOVA: Summary, Assumptions, and the Goal of the analysis
- One-Way ANOVA as a model of Multiple Linear Regression
- One-Way ANOVA: the *F*-test

ANOVA: Introduction

The methods of the **Analysis of Variance** (ANOVA) are special cases of the methods of the **Multiple Linear Regression**.

- In essence, we measure one numerical statistical variable, denoted by “y”, that is a quantitative data item of some statistical data units.
- In addition, we consider one or more (qualitative / quantitative) factors.
- Depending upon the number of the factors, we distinguish:
  - — one-factor = one-way ANOVA
  - — two-factor = two-way ANOVA

ANOVA: Introduction

- We assume that each of the factors can attain only finitely many distinct values.
- Hence, there are only finitely many possible combinations (a Cartesian product) of all possible values of the factors.
- Each combination of the values constitutes one group of the statistical units.
- The purpose is to study whether the expected value $\mathrm{E}[y]$ of the statistical variable "y" depends upon the values of the parameters, or not; that is, whether the expected value in each group is the same, or not.
- ANOVA is based upon the theory of Multiple Linear Regression.
- We shall study one-factor ANOVA in this lecture.

One-factor ANOVA: Motivation: Example 1

Consider a gross sample of n patients who have been cured for some disease.

The patients were divided into k groups:

— There are n_1 patients in the 1^{st} group.

— There are n_2 patients in the 2^{nd} group.

— ...

— There are n_k patients in the k^{th} group.

→ There are $n_1 + n_2 + \cdots + n_k = n$ patients in total.

One-factor ANOVA: Motivation: Example 1

Each of the n_1 patients in the 1st group has been cured by method no. 1.

Each of the n_2 patients in the 2nd group has been cured by method no. 2.

...

Each of the n_k patients in the k^{th} group has been cured by method no. k.

We then measure the success of the medical treatment.

That is, let "y" be a (quantitative) variable meaning the health of the patient.

One-factor ANOVA: Motivation: Example 1

Measuring the health of each of the n_1 patients in the 1^{st} group,

we have the values $y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$

Measuring the health of each of the n_2 patients in the 2^{nd} group,

we have the values $y_{21}, y_{22}, \dots, y_{2n_2}$

...

...

Measuring the health of each of the n_k patients in the k^{th} group,

we have the values $y_{k1}, y_{k2}, y_{k3}, y_{k4}, \dots, y_{kn_k}$

One-factor ANOVA: Motivation: Example 1

Having got the results of the measurements of the health

group no. 1: $y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$
group no. 2: $y_{21}, y_{22}, \dots, y_{2n_2}$
group no. 3: $y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, \dots, y_{3n_3}$
... ...
group no. k: $y_{k1}, y_{k2}, \dots, y_{kn_k}$

we test **the null hypothesis** that:

- there are no differences among the treatments, that is
- the expected values of the health in the groups are approximately the same

One-factor ANOVA: Motivation: Example 2

We test k distinct cars. We test the 1st car n_1 times,

we test the 2nd car n_2 times, etc., and we test the k^{th} car n_k times

for mileage (fuel consumption per 100 km).

Then $y_{11}, y_{12}, \dots, y_{1n_1}$, $y_{21}, y_{22}, \dots, y_{2n_2}$, etc., $y_{k1}, y_{k2}, \dots, y_{kn_k}$

are the results of the measurements, i.e. the fuel consumptions per 100 km.

We test **the null hypothesis** that:

the average fuel consumption of each car is the same.

One-Way ANOVA

- Summary
- Assumptions
- The classical assumptions
- The purpose of the Analysis of Variance

One-way ANOVA: Summary

We have got the sample of the k groups

of the $n = n_1 + n_2 + \cdots + n_k$ observations

$$y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$$
$$y_{21}, y_{22}, \dots, y_{2n_2}$$
$$y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, \dots, y_{3n_3}$$
$$\dots$$
$$y_{k1}, y_{k2}, \dots, y_{kn_k}$$

where $y_{ij} \in \mathbb{R}$ for $j = 1, 2, \dots, n_i$ for $i = 1, 2, \dots, k$.

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

One-way ANOVA: Summary

First:

- A sample of n statistical units was selected from a larger population.
- Each of the statistical units was placed into its respective group $i \in \{1, 2, \dots, k\}$ and measured, so we have obtained the values y_{ij} for $j = 1, 2, \dots, n_i$ for $i = 1, 2, \dots, k$. (Where n_i is the number of the units finally found in the *i*-th group for $i = 1, 2, \dots, k$. In the end, it holds $n = n_1 + n_2 + \dots + n_k$.)
- We assume $y_{ij} \approx \mu_i$ and we have $y_{ij} = \mu_i + \varepsilon_{ij}$, where ε_{ij} is a <u>random deviation</u> (error).
- The random deviation is caused by the intrinsic properties of the statistical unit

One-way ANOVA: Summary

Second:

- We prepare the groups $i = 1, 2, \dots, k$ at the beginning.
- When measuring the value y_{ij}, we select a unit from the group i first

  and measure its value y_{ij} then for $j = 1, 2, \dots, n_i$ for $i = 1, 2, \dots, k$.
- The random deviation ε_i here is caused

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

  unconsidered factors),

  or by random errors in the measurement itself. (!)

One-way ANOVA: Summary

Remarks:

- In practice, the data may be obtained in either way (first or second).
- In either case (first or second), the group which the unit belongs to is assumed to be known exactly, i.e. without any doubts.
- Assuming $y_{ij} \approx \mu_i$, even the dependent values y_{ij} may be measured exactly, i.e. without any measurement error, the random deviation $\varepsilon_{ij} = y_{ij} - \mu_i$ being caused by the intrinsic properties (other unknown / "random" / unconsidered factors).
- For the purpose of the mathematical analysis, we assume the second case only.

One-way ANOVA: Assumptions

- There are k groups of sizes $n_1, n_2, \dots, n_k$ given, known, and fixed before the measurements.
- Moreover, we are given k numbers $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}$, which are the expected values of the variable "y" in the respective groups.
- We have $n = n_1 + n_2 + \cdots + n_k$ random variables

$$Y_{11}, Y_{12}, \dots, Y_{1n_1}, \quad Y_{21}, Y_{22}, \dots, Y_{2n_2}, \quad \dots \dots \dots, \quad Y_{k1}, Y_{k2}, \dots, Y_{kn_k}$$

which are assumed to be independent (or uncorrelated).
- We also have n random variables

$$\varepsilon_{11}, \varepsilon_{12}, \dots, \varepsilon_{1n_1}, \quad \varepsilon_{21}, \varepsilon_{22}, \dots, \varepsilon_{2n_2}, \quad \dots \dots \dots, \quad \varepsilon_{k1}, \varepsilon_{k2}, \dots, \varepsilon_{kn_k}$$

One-way ANOVA: Assumptions

- We stack the group expected values $\mu_1, \mu_2, \dots, \mu_k$ into a vector:

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \vdots \\ \vdots \\ \mu_k \\ \vdots \\ \mu_k \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_1 - \text{times} \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_2 - \text{times} \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} \dots - \text{times} \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_k - \text{times} \end{matrix}$$

One-way ANOVA: Assumptions

- We stack the random variables $Y_{11},\dots,Y_{1n_1},\ Y_{21},\dots,Y_{2n_2},\ \dots\dots\dots,\ Y_{k1},\dots,Y_{kn_k}$ into a random vector:

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \vdots \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{—times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2\text{—times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \dots\text{—times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_k\text{—times} \end{array}$$

One-way ANOVA: Assumptions

- We stack the random variables $\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \ \varepsilon_{21}, \dots, \varepsilon_{2n_2}, \ \dots\dots\dots, \ \varepsilon_{k1}, \dots, \varepsilon_{kn_k}$ into a random vector:

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_1 - \text{times} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_2 - \text{times} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \dots - \text{times} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} n_k - \text{times} \end{matrix}$$

One-way ANOVA: The Classical Assumptions

- We have the underlying probability space $(\Omega, \mathcal{F}, P)$.
- Let $\omega \in \Omega$ be the outcome of the random experiment.
- We have

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

<u>In other words:</u>

- The measured values $y_{11}, \dots, y_{1n_1},\ y_{21}, \dots, y_{2n_2},\ \dots,\ y_{k1}, \dots, y_{kn_k}$ are the numerical outcomes $Y_{11}(\omega), \dots, Y_{kn_k}(\omega)$ of the random experiment.
- The numerical outcomes $Y_{11}(\omega), \dots, Y_{kn_k}(\omega)$ are obtained so that the numerical outcomes $\varepsilon_{11}(\omega), \dots, \varepsilon_{kn_k}(\omega)$ of the random experiment

One-way ANOVA: The Classical Assumptions

The classical assumptions of One-Way ANOVA are:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

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

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

<u>It follows that:</u>

$$\mathrm{E}[Y_{ij}] = \mu_i \qquad \text{and} \qquad \mathrm{E}[\varepsilon_{ij}] = 0$$

One-way ANOVA: The Classical Assumptions

The classical assumptions $Y \sim \mathcal{N}(\mu, \sigma^2 I)$ 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,

- $\mathrm{Var}(Y_{ij}) = \mathrm{Var}(\varepsilon_{ij}) = \sigma^2$ for $j = 1, 2, \dots, n_i$
  for $i = 1, 2, \dots, k$ for some $\sigma^2 \in \mathbb{R}$
- and the random variables $Y_{11}, \dots, Y_{kn_k}$ or $\varepsilon_1, \dots, \varepsilon_{kn_k}$

homoscedasticity, i.e. the variance is the same

One-way ANOVA: The purpose of the analysis

By the Classical Assumptions, we have:

$$\mathrm{E}[Y_{11}] = \mathrm{E}[Y_{12}] = \cdots = \mathrm{E}\left[Y_{1n_1}\right] = \mu_1$$

$$\mathrm{E}[Y_{21}] = \mathrm{E}[Y_{22}] = \cdots = \mathrm{E}\left[Y_{2n_2}\right] = \mu_2$$

$$\cdots$$

$$\mathrm{E}[Y_{k1}] = \mathrm{E}[Y_{k2}] = \cdots = \mathrm{E}\left[y_{kn_k}\right] = \mu_k$$

The purpose is to test the null hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

One-way ANOVA: Solution

Given the sample

$$y_{11}, y_{12}, y_{13}, \dots, y_{1n_1}$$
$$y_{21}, y_{22}, \dots, y_{2n_2}$$
$$y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, \dots, y_{3n_3}$$
$$\dots$$
$$y_{k1}, y_{k2}, \dots, y_{kn_k}$$

of the random variables

$$Y_{11}, Y_{12}, Y_{13}, \dots, Y_{1n_1}$$
$$Y_{21}, Y_{22}, \dots, Y_{2n_2}$$
$$Y_{31}, Y_{32}, Y_{33}, Y_{34}, Y_{35}, \dots, Y_{3n_3}$$
$$\dots$$
$$Y_{k1}, Y_{k2}, \dots, Y_{kn_k}$$

that is

One-way ANOVA: Solution

…and
assuming (among others) that the expected values are the same in each group

$$\mathrm{E}[Y_{ij}] = \mu_i \qquad \text{for} \quad j = 1, 2, \ldots, n_i \qquad \text{for} \quad i = 1, 2, \ldots, k$$

it is our purpose to test the null hypothesis that

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

<u>SOLUTION:</u>

- We shall express this situation as a model of Multiple Linear Regression.
- We shall then use the theory of Multiple Linear Regression ("Theorem 8") to make up an *F*-test for the null hypothesis H_0.

One-Way ANOVA as a model of Multiple Linear Regression

One-way ANOVA: Model of Multiple Linear Regression

Given the random variables $Y_{11}, Y_{12}, \dots, Y_{1n_1}, Y_{21}, Y_{22} \dots, Y_{2n_2}, \dots\dots\dots, Y_{k1}, Y_{k2}, \dots, Y_{kn_k},$

we can express the original assumption (on the left)

in terms of Multiple Linear Regression (on the right):

$$
\begin{array}{lcll}
\mathrm{E}[Y_{1j}] = \mu_1 & \rightarrow & \mathrm{E}[Y_{1j}] = \mu_1 & \text{for} \quad j = 1, 2, \dots, n_1 \\
\mathrm{E}[Y_{2j}] = \mu_2 & \rightarrow & \mathrm{E}[Y_{2j}] = \mu_1 + \beta_2 & \text{for} \quad j = 1, 2, \dots, n_2 \\
\mathrm{E}[Y_{3j}] = \mu_3 & \rightarrow & \mathrm{E}[Y_{3j}] = \mu_1 \qquad + \beta_3 & \text{for} \quad j = 1, 2, \dots, n_3 \\
\vdots & & \vdots \qquad\qquad \ddots & \\
\mathrm{E}[Y_{kj}] = \mu_k & \rightarrow & \mathrm{E}[Y_{kj}] = \mu_1 \qquad\qquad + \beta_k & \text{for} \quad j = 1, 2, \dots, n_k
\end{array}
$$

One-way ANOVA: Model of Multiple Linear Regression

Having

$$\mathrm{E}\left[Y_{1j}\right] = \mu_1 \qquad \text{for} \quad j = 1, 2, \dots, n_1$$

$$\mathrm{E}\left[Y_{ij}\right] = \mu_i = \mu_1 + \beta_j \qquad \text{for} \quad j = 1, 2, \dots, n_i \qquad \text{for} \quad i = 2, \dots, k$$

the original hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

is equivalent with

$$H_0: \quad \beta_2 = \cdots = \beta_k = 0$$

which we can test (recall "Theorem 8").

One-way ANOVA: Model of Multiple Linear Regression

Putting yet $\beta_1 = \mu_1$ and by using the notation

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \vdots \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{–times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2\text{–times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \ldots\text{–times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_k\text{–times} \end{matrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{–times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2\text{–times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \ldots\text{–times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_k\text{–times} \end{matrix}$$

One-way ANOVA: Model of Multiple Linear Regression

… and by considering the matrix

$$
\boldsymbol{X} = \begin{array}{c}
\begin{array}{cccc} \overbrace{}^{\beta_1} & \overbrace{}^{\beta_2} & \overbrace{}^{\dots} & \overbrace{}^{\beta_k} \end{array} \\
\begin{pmatrix}
1 & & & \\
\vdots & & & \\
1 & & & \\
1 & 1 & & \\
\vdots & \vdots & & \\
1 & 1 & & \\
\vdots & & & \\
\vdots & & \ddots & \\
\vdots & & & \\
1 & & & 1 \\
\vdots & & & \vdots \\
1 & & & 1
\end{pmatrix}
\end{array}
\begin{array}{l}
\left.\begin{array}{c} \\ \\ \\ \end{array}\right\} n_1\text{—times} \\
\left.\begin{array}{c} \\ \\ \\ \end{array}\right\} n_2\text{—times} \\
\left.\begin{array}{c} \\ \\ \\ \end{array}\right\} \dots\text{—times} \\
\left.\begin{array}{c} \\ \\ \\ \end{array}\right\} n_k\text{—times}
\end{array}
$$

Notice that

$$\operatorname{rank}(\boldsymbol{X}) = k$$

One-way ANOVA: Model of Multiple Linear Regression

… we can write the One-Way ANOVA in terms of Multiple Linear Regression as follows:

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

where $\beta_1 = \mu_1$ plays the rôle of the intercept term

and $\beta_2 = \mu_2 - \mu_1$, …, $\beta_k = \mu_k - \mu_1$ are the other regression coefficients.

Recall that we wish to test the null hypothesis

$$H_0: \quad \beta_2 = \cdots = \beta_k = 0$$

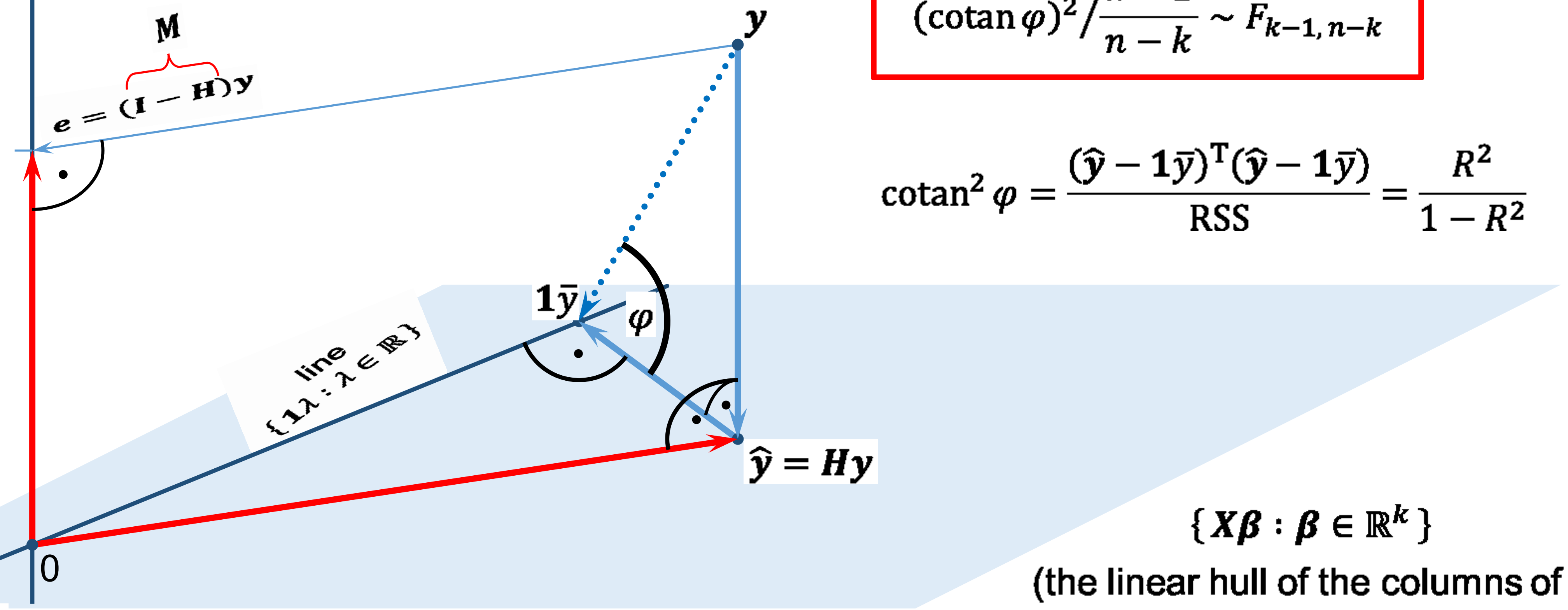
The Coefficient of Determination (R^2): Th. 8: Corollary

Let: $\boxed{\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}$

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

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

$\boxed{n-k}$



By Theorem 8:

$$\boxed{(\cot\varphi)^2 \Big/ \frac{k-1}{n-k} \sim F_{k-1,\,n-k}}$$

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

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

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

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

the line is a subspace of dimension $\boxed{1}$

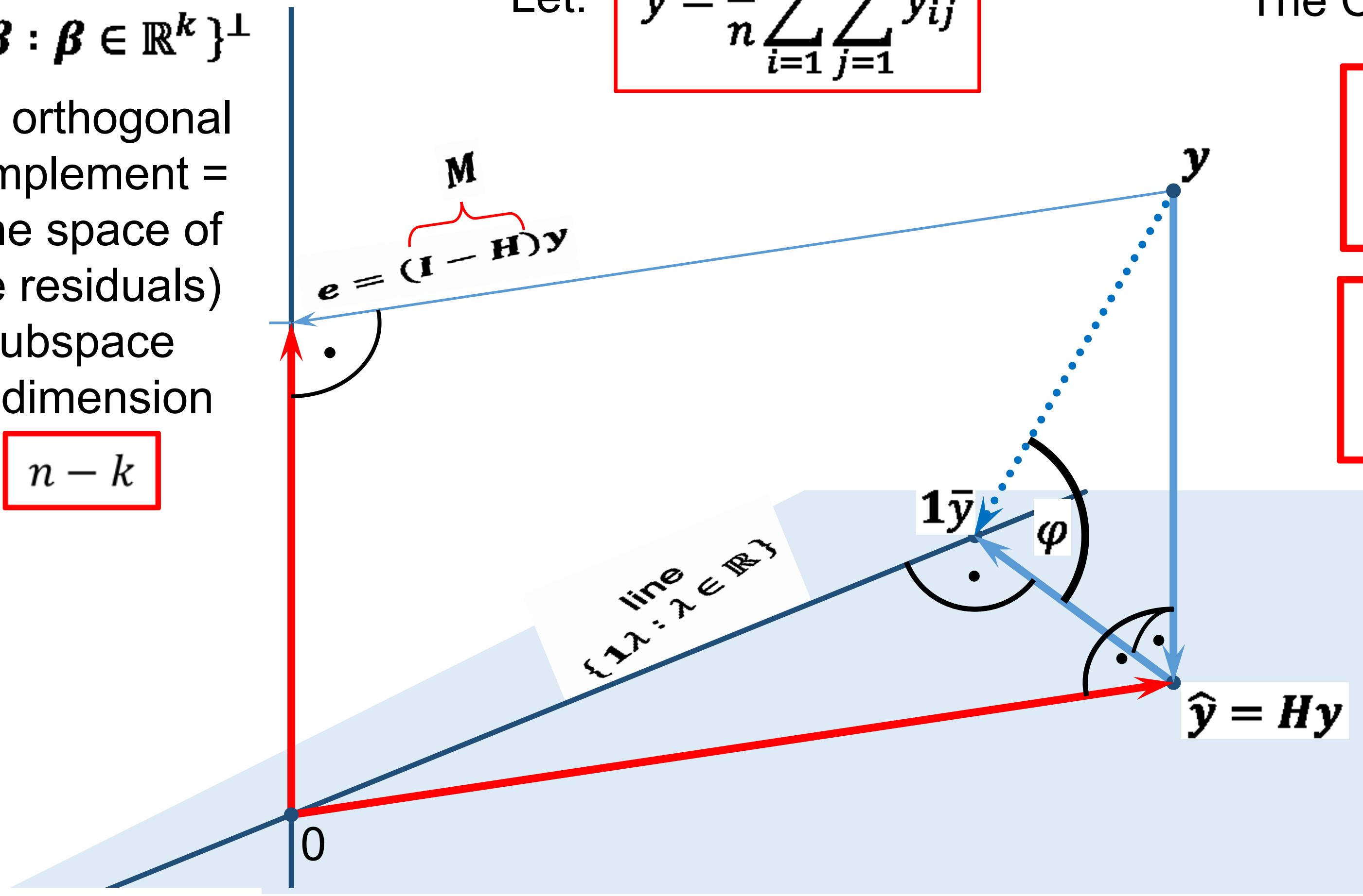
the dimension of its complement within the subspace of dimension k is $\boxed{\ \ -1}$

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

Let: $\boxed{\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}$

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

(the orthogonal complement = = the space of the residuals) subspace of dimension $\boxed{n-k}$



The Coefficient of Determination:

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

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

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

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

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

the line is a subspace of dimension $\boxed{1}$

the dimension of its complement within the subspace of dimension k is $\boxed{-1}$

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

Let: $\boxed{\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}$

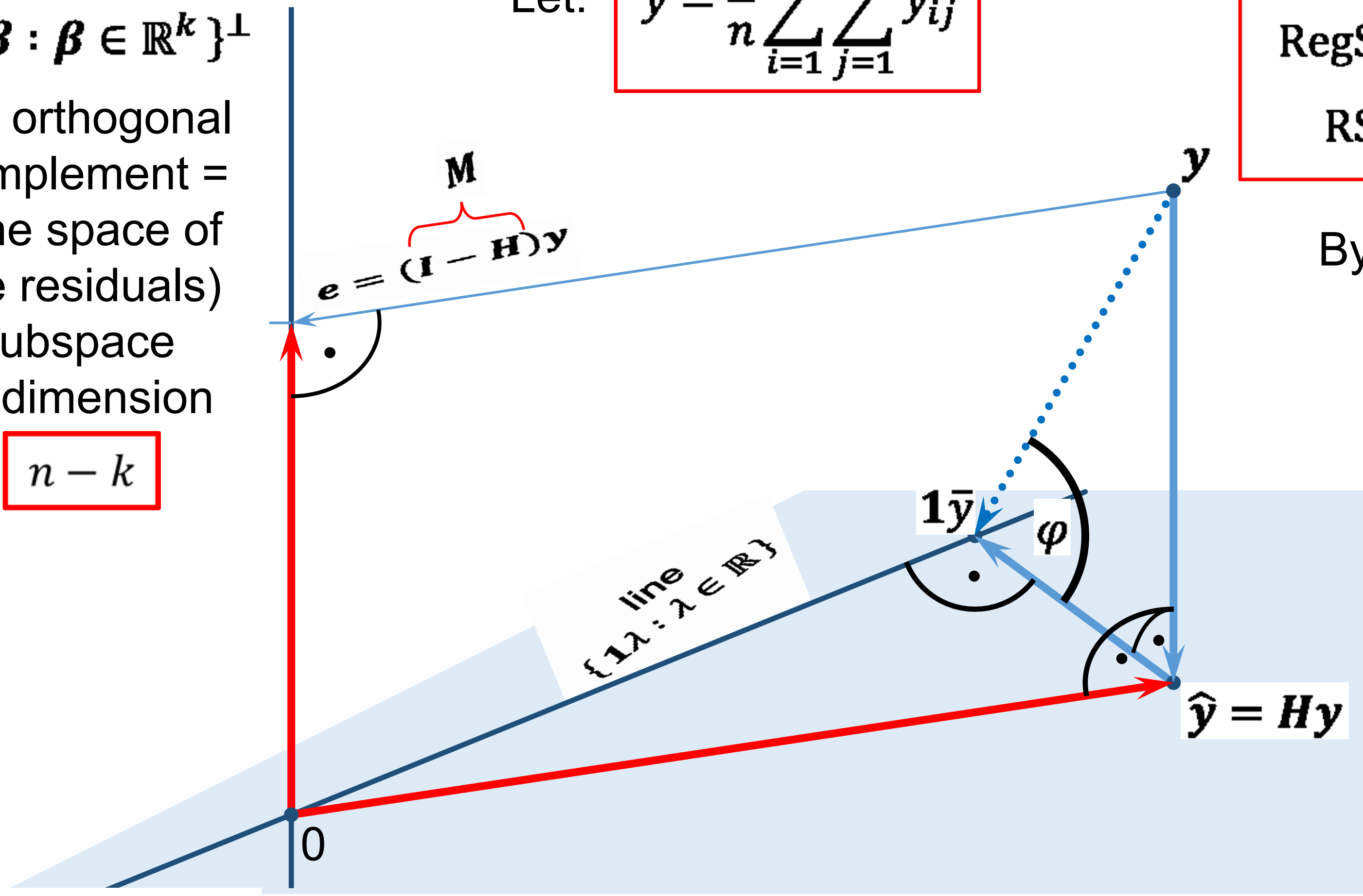
$$\boxed{\begin{aligned} \mathrm{TSS} &= (\boldsymbol{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\bar{y}) \\ \mathrm{RegSS} &= (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) \\ \mathrm{RSS} &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y}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

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

(the orthogonal complement = = the space of the residuals) subspace of dimension $\boxed{n-k}$

By the Pythagoras Theorem:

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



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

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

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

the line is a subspace of dimension $\boxed{1}$

the dimension of its complement within the subspace of dimension k is $\boxed{-1}$

The Coefficient of Determination (R^2)

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

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

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

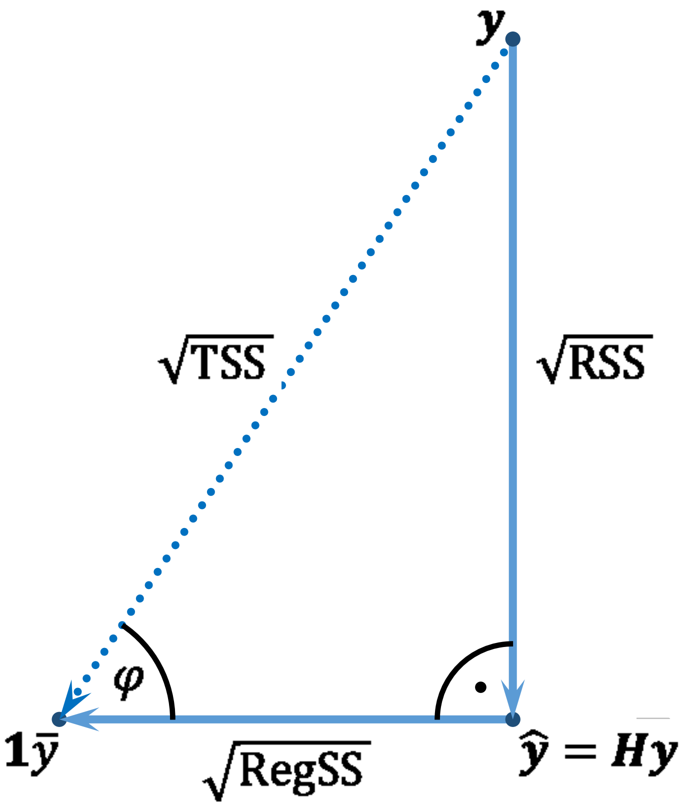
$$\mathrm{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$$

Coefficient of Determination:

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

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

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



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

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

$$\beta_2 = \cdots = \beta_k = 0$$

it holds

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

$$= \frac{\mathrm{RegSS}}{\mathrm{RSS}} \Big/ \frac{k-1}{n-k} \sim F_{k-1,\,n-k}$$

One-way ANOVA: Model of Multiple Linear Regression

We have

$$\mathrm{RegSS} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{\boldsymbol{y}} - \mathbf{1}\bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{y}_{ij} - \bar{y})^2 \qquad \text{where} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

where

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

where

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

where

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

One-way ANOVA: Model of Multiple Linear Regression

Calculate:

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \begin{pmatrix} n & n_2 & n_3 & n_4 & \cdots & n_k \\ n_2 & n_2 & & & & \\ n_3 & & n_3 & & & \\ n_4 & & & n_4 & & \\ \vdots & & & & \ddots & \\ n_k & & & & & n_k \end{pmatrix}$$

where

$$n = n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i$$

One-way ANOVA: Model of Multiple Linear Regression

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

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = \begin{pmatrix} +{}^{1}/_{n_1} & -{}^{1}/_{n_1} & -{}^{1}/_{n_1} & -{}^{1}/_{n_1} & \cdots & -{}^{1}/_{n_1} \\ -{}^{1}/_{n_1} & {}^{1}/_{n_2} + {}^{1}/_{n_1} & +{}^{1}/_{n_1} & +{}^{1}/_{n_1} & \cdots & +{}^{1}/_{n_1} \\ -{}^{1}/_{n_1} & +{}^{1}/_{n_1} & {}^{1}/_{n_3} + {}^{1}/_{n_1} & +{}^{1}/_{n_1} & \cdots & +{}^{1}/_{n_1} \\ -{}^{1}/_{n_1} & +{}^{1}/_{n_1} & +{}^{1}/_{n_1} & {}^{1}/_{n_4} + {}^{1}/_{n_1} & \cdots & +{}^{1}/_{n_1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -{}^{1}/_{n_1} & +{}^{1}/_{n_1} & +{}^{1}/_{n_1} & +{}^{1}/_{n_1} & \cdots & {}^{1}/_{n_k} + {}^{1}/_{n_1} \end{pmatrix}$$

One-way ANOVA: Model of Multiple Linear Regression

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

$$
\begin{array}{rcl}
 & \begin{array}{cccc} \overbrace{\qquad\qquad\qquad}^{n_1} & \overbrace{\qquad\qquad\qquad}^{n_2} & \overbrace{\qquad\qquad}^{\cdots} & \overbrace{\qquad\qquad\qquad}^{n_k} \end{array} & \\
\boldsymbol{H} = & \begin{pmatrix}
\begin{matrix} 1/n_1 & \cdots & 1/n_1 \\ \vdots & \ddots & \vdots \\ 1/n_1 & \cdots & 1/n_1 \end{matrix} & & & \\
 & \begin{matrix} 1/n_2 & \cdots & 1/n_2 \\ \vdots & \ddots & \vdots \\ 1/n_2 & \cdots & 1/n_2 \end{matrix} & & \\
 & & \begin{matrix} \ddots & \cdots & \ddots \\ \vdots & \ddots & \vdots \\ \ddots & \cdots & \ddots \end{matrix} & \\
 & & & \begin{matrix} 1/n_k & \cdots & 1/n_k \\ \vdots & \ddots & \vdots \\ 1/n_k & \cdots & 1/n_k \end{matrix}
\end{pmatrix} &
\begin{array}{l}
\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{—times} \\
\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2\text{—times} \\
\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \ldots\text{—times} \\
\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_k\text{—times}
\end{array}
\end{array}
$$

One-way ANOVA: Model of Multiple Linear Regression

Calculate:

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

That is:

$$\hat{y}_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \text{for} \quad i = 1, 2, \dots, k$$

Denote:

$$\bar{y}_i = \hat{y}_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \text{for} \quad i = 1, 2, \dots, k$$

and call this ↗ quantity **the sample mean of the i-th group** for $i = 1,2, \dots, k$.

One-way ANOVA: Model of Multiple Linear Regression

Calculate:

$$\text{RegSS} = (\hat{y} - \mathbf{1}\bar{y})^{\mathrm{T}}(\hat{y} - \mathbf{1}\bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{y}_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2$$

Denote:

$$\text{SS}_{\text{B}} = \text{RegSS} = \sum_{i=1}^{k} n_i \times (\bar{y}_i - \bar{y})^2$$

and call this ↗ quantity **the sum of squares "between"**

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$

One-way ANOVA: Model of Multiple Linear Regression

Calculate:

$$\mathrm{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

Denote:

$$\mathrm{SS_W} = \mathrm{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

and call this ↗ quantity **the sum of squares "within"**

where $$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

One-way ANOVA: Remarks: Sample variances

The quantity

$$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{\mathrm{DF_B}} = \frac{\mathrm{RegSS}}{k-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2}{k-1}$$

is the sample variance between the groups.

The quantity

$$\mathrm{MS_W} = \frac{\mathrm{SS_W}}{\mathrm{DF_W}} = \frac{\mathrm{RSS}}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2}{n-k}$$

is the sample variance within the groups.

One-Way ANOVA: The *F*-test

- The *F*-test

One-Way ANOVA: the *F*-test

To test the null hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

against the alternative hypothesis

$$H_1: \quad \mu_i \neq \mu_j \qquad \text{for some} \quad i, j \in \{1, 2, \dots, k\}$$

calculate the statistic

$$F = \frac{\mathrm{SS_B}}{\mathrm{SS_W}} \Big/ \frac{\mathrm{DF_B}}{\mathrm{DF_W}} = \frac{\mathrm{RegSS}}{\mathrm{RSS}} \Big/ \frac{k-1}{n-k} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \Big/ \frac{k-1}{n-k}$$

One-Way ANOVA: the *F*-test

Recall that, if H_0 holds true, then

$$F \sim F_{k-1,n-k}$$

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

and calculate the **critical value**

$$c = F_{k-1,n-k}(1-\alpha)$$

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.