



**COMILLAS**  
UNIVERSIDAD PONTIFICIA

**ICAI**

ICAI – GITI/GITT

# Presentación de la asignatura Estadística II

Eugenio Sánchez Úbeda

January 2024

**comillas.edu**





COMILLAS  
UNIVERSIDAD PONTIFICIA

ICAI ICADE GHS

# Información general




# Información general

## Estadística II: La base del iceberg



# Business Analytics Spectrum

- **Descriptive:** statistics (data analysis, analysis of variance, correlation)
- **Predictive:** simulation, regression, forecasting
- **Prescriptive:** optimization, heuristics, decision analysis



Stochastic Optimization	How can we achieve the best outcome including the effects of variability?	<b>PRESCRIPTIVE</b>
Optimization	How can we achieve the best outcome?	
Predictive modeling	What will happen next if?	<b>PREDICTIVE</b>
Forecasting	What if these trends continue?	
Simulation	What could happen...?	
Alerts	What actions are needed?	
Query/drill down	What exactly is the problem?	<b>DESCRIPTIVE</b>
Ad hoc reporting	How many, how often, where?	
Standard reporting	What happened?	

Source: A. Fleischer et al. *ILOG Optimization for Collateral Management*

# Información general

## Objetivos del Aprendizaje

- Bloque 1
  - Comprensión de conceptos teóricos
  - Aplicación de los conceptos a la resolución de problemas prácticos
  - Análisis e interpretación de los resultados
- Bloque 2
  - Uso de software estadístico



Profesor	Grupo-Aula	correo	Despacho
Juan Pablo Fuentes	3º A – 203 3º D - 408	jpfuentes@icai.comillas.edu	
Eugenio Sánchez (coordinador)	3º B – 309	eugenio.sanchez@comillas.edu	SM26.D-304
Andrés Ramos	3º C – 304	andres.ramos@comillas.edu	SM26.D-103
Álvaro Guinea Juliá	3º E+F – 213	agjulia@icai.comillas.edu	507
Anne Coll	Apoyo prácticas 3º B y E+F	amcoll@comillas.edu	

# Información general

## Organización

- Calendario académico (14 semanas – 13 reales)

### 2024

Enero								Sem.	Febrero								Sem.	Marzo								Sem.	Abril								Sem.
L	M	X	J	V	S	D		L	M	X	J	V	S	D		L	M	X	J	V	S	D		L	M	X	J	V	S	D					
1	2	3	4	5	6	7					1	2	3	4	3					1	2	3	7	1	2	3	4	5	6	7	11				
8	9	10	11	12	13	14		5	6	7	8	9	10	11	4	4	5	6	7	8	9	10	8	8	9	10	11	12	13	14	12				
15	16	17	18	19	20	21	1	12	13	14	15	16	17	18	5	11	12	13	14	15	16	17	9	15	16	17	18	19	20	21	13				
22	23	24	25	26	27	28	2	19	20	21	22	23	24	25	6	18	19	20	21	22	23	24	10	22	23	24	25	26	27	28	14				
29	30	31					3	26	27	28	29				7	25	26	27	28	29	30	31		29	30										

- Normas
  - Puntualidad británica
  - Prohibido uso inapropiado de móviles y demás aparatos
- SIFO-Moodle
  - Almacén de material
  - Realización de las prácticas

INTRODUCCION
1. PRESENTACIÓN DE LA ASIGNATURA
2. P00: REPASO MATLAB
3. T1: REGRESIÓN LINEAL
4. P01: PRÁCTICA REGRESIÓN LINEAL
5. T2: CLASIFICACIÓN

# Información general

## Organización

- Tipos de sesiones
  - **Clases teóricas** (con enfoque práctico)
    - Tipo de problema a resolver
    - Modelado estadístico del problema
    - Interpretación del modelo
    - Ejemplos de aplicación
  - **Prácticas en laboratorio**
    - 5 sesiones de 2 horas contiguas
    - Herramienta Matlab
  - **Pruebas intermedias**
    - 2 test conceptos
    - Realimentación en ambos sentidos





2024

- Distribución prevista de los temas

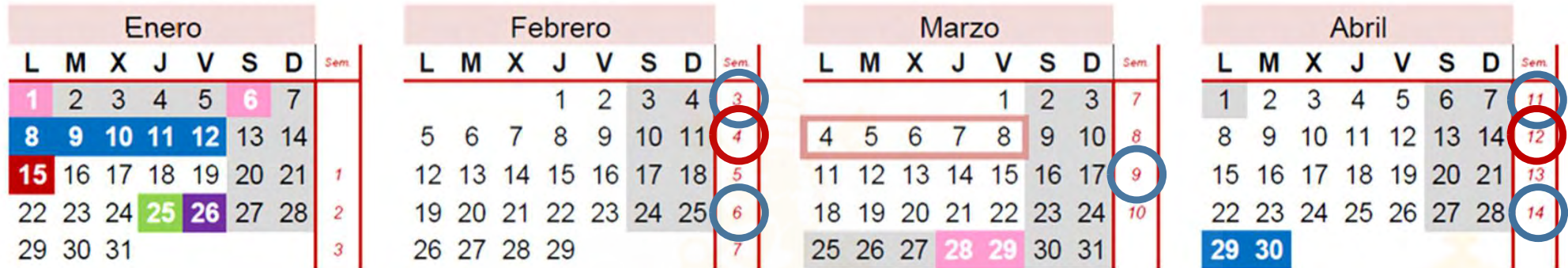


Tema	Semana prevista
T0: Introducción y repaso conceptos	S1
T1: Regresión lineal	S1-S3
T2: Clasificación	S4-S6
T3: Análisis de la varianza	S7-S9
T4: Análisis de componentes principales	S9-S11
T5: Análisis factorial	S9-S11
T6: Análisis de conglomerados	S12-S14

# Información general

## Organización

- Hitos



Prácticas (sesión 2 horas)	Semana prevista
P1: Práctica Regresión lineal	S3
P2: Práctica Clasificación	S6
P3: Práctica ANOVA	S9
P4: Práctica PCA/FA	S11
P5: Práctica Clustering	S14
Pruebas	Semana prevista
Prueba 1 (Regresión)	S4
Prueba 2 (ANOVA y PCA)	S12

# Información general

## Organización

- **Prácticas**

- Duran dos horas (100 minutos)
- Cuentan con un profesor de apoyo para resolver dudas
- Al final de la práctica hay un test sobre la misma

Grupo	Día (2 horas)	Profesor	Profesor de apoyo
A	Viernes	Juan Pablo	Andrés
B	Martes	Eugenio	Anne
C	Jueves	Andrés	Juan Pablo
D	Martes	Juan Pablo	Álvaro
E-F	Miércoles	Álvaro	Anne

# Información general

## Material

- Apuntes
  - Presentaciones utilizadas en clase
  - Enunciados y código para las prácticas

- Libros de referencia



- G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, An Introduction to Statistical Learning with Applications in Python, Springer, 2023 (<https://www.statlearning.com>)

- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Ed., Springer, 2017  
([https://hastie.su.domains/ElemStatLearn/printings/ESLII\\_print12.pdf](https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12.pdf))



# Información general

## Material de referencia

- G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, An Introduction to **Statistical Learning** with **Applications in Python**, Springer, 2023

### 3.2.1 Estimating the Regression

As was the case in the simple linear regression, the coefficients  $\beta_0, \beta_1, \dots, \beta_p$  in (3.19) are unknown. To obtain estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ , we can make

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$$

The parameters are estimated using the method of least squares, which we saw in the context of simple linear regression to minimize the sum of squared residuals:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

The values  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  that minimize the RSS are the regression coefficient estimates. In general, the formulas for these estimates given in (3.4), the multivariate normal distribution, are somewhat complicated forms that are not amenable to matrix algebra. For this reason, we do not provide them here. Instead, a software package can be used to compute these coefficient estimates, and later in this chapter we will show how this can be done in R. Figure 3.4

We extract the response, and fit the model.

```
In [10]: y = Boston['medv']
         model = sm.OLS(y, X)
         results = model.fit()
```

Note that `sm.OLS()` does not fit the model; it specifies the model, and then `model.fit()` does the actual fitting.

Our ISLP function `summarize()` produces a simple table of the parameter estimates, their standard errors, t-statistics and p-values. The function takes a single argument, such as the object `results` returned here by the `fit` method, and returns such a summary.

```
In [11]: summarize(results)
```

```
Out [11]:
```

	coef	std err	t	P> t
intercept	34.5538	0.563	61.415	0.0
lstat	-0.9500	0.039	-24.528	0.0

Before we describe other methods for working with fitted models, we outline a more useful and general framework for constructing a model matrix  $X$ .



# Información general

## Material complementario

- Peña, D., Análisis de datos multivariantes. Ed. McGraw-Hill. Madrid. 2002.
- Cho, M., and Martinez, W. L. (2014). Statistics in MATLAB: A Primer. Chapman & Hall/CRC Computer Science & Data Analysis.
- Graffelman, J. (2011). Multivariate Analysis with Matlab and R. Chapman & Hall, CRC Press.
- Maté, C. (1995). Curso General sobre Statgraphics. Procedimientos. Métodos Estadísticos. Aplicaciones. Ejercicios Resueltos. Tomo II. Editorial Universidad Pontificia Comillas. Madrid.
- Montgomery, D.C. and Runger, G.C. (2002). Probabilidad y Estadística Aplicadas a la Ingeniería. Limusa Wiley. México D.F.
- Johnson, R. A. and Wichern, D. W. (1998). Applied Multivariate Statistical Analysis. Fourth Edition. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.
- Mickey, R. M.; Dun, O.J. and Clark, V. A. (2004). Applied Statistics. Analysis of Variance and Regression. Third Edition, John Wiley & Sons. New York.
- Morrison, D. F. (1990). Multivariate Statistical Methods. Third Edition, McGraw-Hill. New York.
- Rencher, A. C. (2002). Methods of Multivariate Analysis. Second Edition, John Wiley & Sons. New York.
- Sharma, S. (1995). Applied Multivariate Techniques. John Wiley & Sons. New York.
- Tabachnick, B. G.; and Fidell, L. S. (1996). Using Multivariate Statistics. Third Edition. HarperCollins College Publishers. New York.
- Tatsuoka, M. M. (1988). Multivariate Analysis. Second Edition, Macmillan Publishing Company. New York.

# Información general

## Tutorías

- Bajo demanda
- Tipos
  - Grupales
    - Objetivo: Aclarar las dudas de los alumnos, no es una clase más en la que se explica temario nuevo
  - Particulares
    - Objetivo: Aclarar dudas demasiado específicas, no respondidas en las tutorías grupales

# Información general

## Idioma

- Inglés y castellano
  - Material en **inglés**, clases en **castellano**
  - Exámenes en **castellano**
- ¿Por qué?
  - Fomenta el **uso de la jerga técnica** en ambos idiomas
    - “Fit a model” / “Ajustar un modelo”
    - “Prune a classification tree” / “Podar un árbol de clasificación”
  - Toda la **documentación técnica de referencia está en inglés** (libros, ayuda de software, blogs, etc.)
  - Utilizando el **castellano en clase y exámenes** se evitan posibles **problemas de comunicación**

# Información general Software

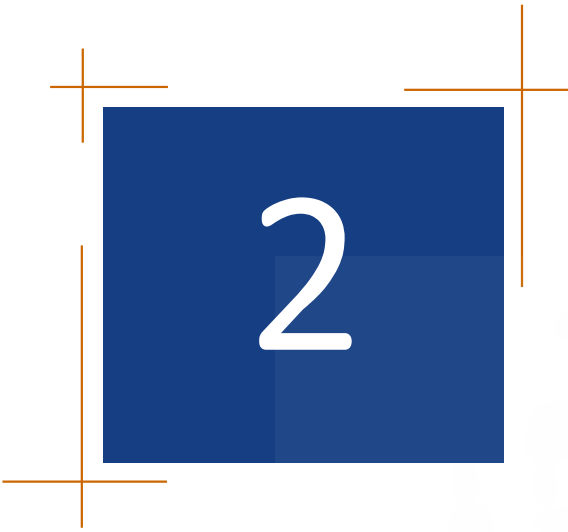
- Lenguaje **Matlab**
  - Utilizado por el profesorado en la preparación del material de clase
  - Utilizado en las prácticas
- ¿Por qué?
  - Se tiene cierta **experiencia**
  - Muy **bien documentado**
  - **Similar a otros lenguajes** interpretados muy utilizados como **R o Python**
  - La **licencia Campus** permite a los alumnos el uso con todo su potencial
  - Se utiliza en otras **disciplinas y asignaturas**

# Información general Software

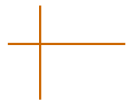


- Necesario tener instalado **Matlab en el portátil propio** para las prácticas
  - Versión **R2020b** o superior
  - Se instala desde [Mathworks.com](https://www.mathworks.com), es necesario crearse una cuenta con la cuenta de correo [@comillas.edu](mailto:@comillas.edu) (licencia Campus)
  - Para la asignatura solo es necesario instalar **Matlab y las toolboxes “Statistics and Machine Learning” y “Deep Learning”**
- También se puede usar **Matlab Online**
  - Disponible en [Mathworks.com](https://www.mathworks.com)
  - Es necesario crearse previamente una cuenta con el correo [@comillas.edu](mailto:@comillas.edu)
- Se recomienda la primera opción

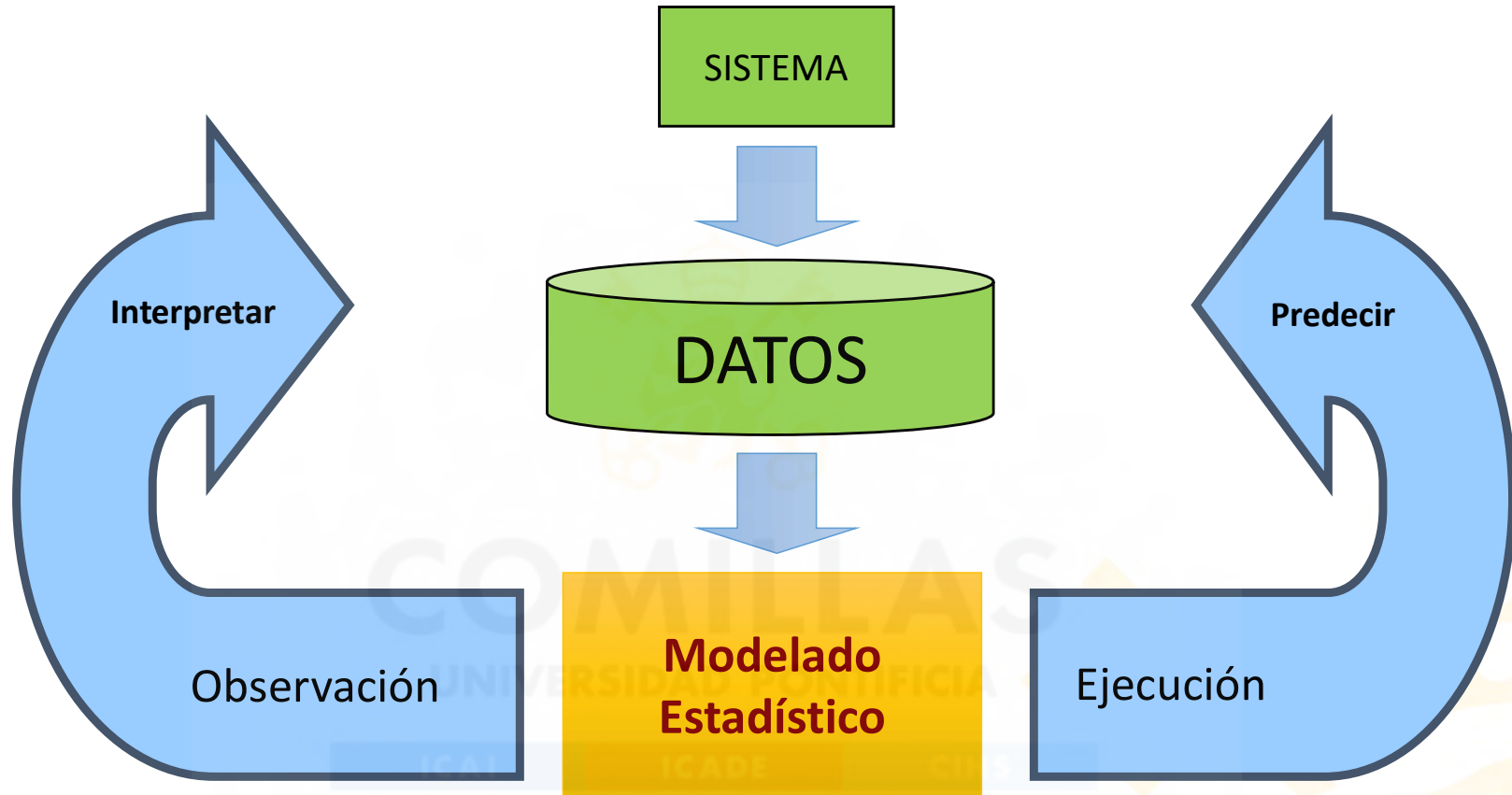




# Motivación



# Motivación Idea



- A partir del sistema se obtienen un conjunto de datos
- Los datos se modelan con el propósito de:
  - interpretar el sistema (por mera observación)
  - predecir el comportamiento del sistema (ejecución del modelo)

# Motivación Ejemplos



During a Formula 1 race, a car sends hundreds of millions of data points to its garage for real-time analysis and feedback. So why not use this detailed and rigorous data system elsewhere, like ... at children's hospitals? Peter van Manen tells us more

Peter van Manen is **Managing Director of McLaren Electronics**

[https://www.ted.com/talks/peter van manen how can formula 1 racing help babies](https://www.ted.com/talks/peter_van_manen_how_can_formula_1_racing_help_babies)

# Motivación Ejemplos



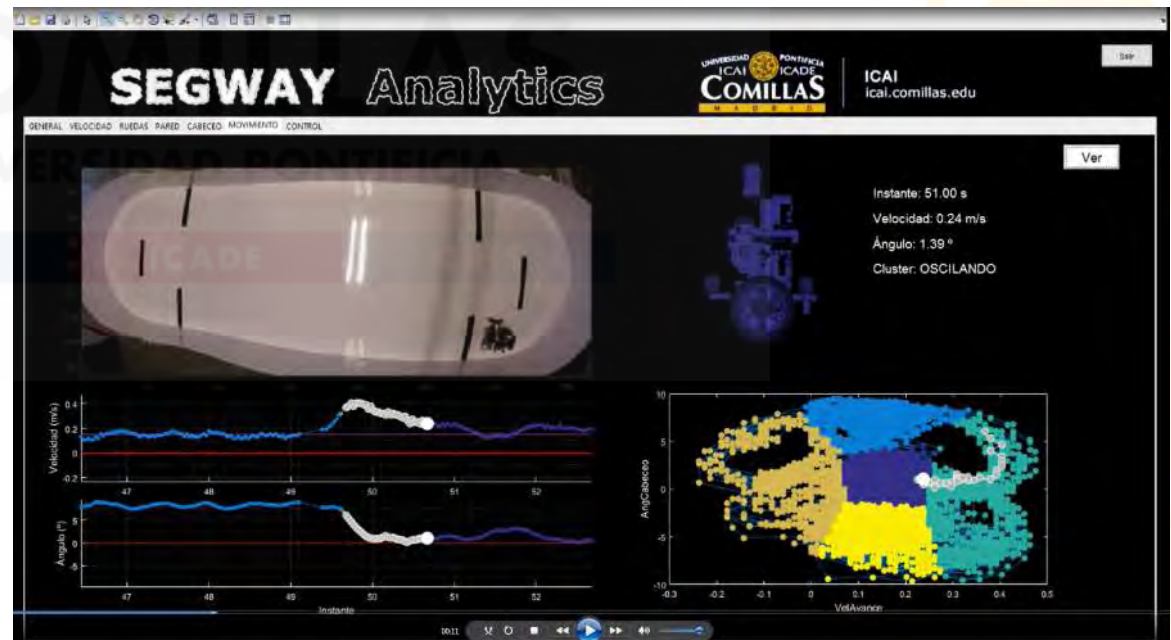
MIT researcher Deb Roy wanted to understand how his infant son learned language -- so he wired up his house with videocameras to catch every moment (with exceptions) of his son's life, then parsed 90,000 hours of home video to watch "gaaaa" slowly turn into "water." Astonishing, data-rich research with deep implications for how we learn.

[http://www.ted.com/talks/deb\\_roy\\_the\\_birth\\_of\\_a\\_word.html](http://www.ted.com/talks/deb_roy_the_birth_of_a_word.html)

# Motivación Ejemplos



Segway construido con Lego





3

COMILLAS  
UNIVERSIDAD PONTIFICIA

## Sistema de evaluación

# Sistema de evaluación

## Convocatoria ordinaria

- La nota se obtiene como:
- Si  $EOR \geq 4.0$

$$N_{OR} = 0.15 \cdot PRA + 0.15 \cdot PRU + 0.15 \cdot EIN + 0.5 \cdot EOR + 0.05 \cdot SCO$$

- Si  $EOR < 4.0$

$$N_{OR} = EOR$$

- Siendo:
  - $EIN$ : Examen intersemestral
  - $EOR$ : Examen ordinario
  - $PRA$ : Prácticas laboratorio con Matlab (media 4 mejores)
  - $PRU$ : Pruebas cortas clase (media)
  - $SCO$ : Participación activa en clase

# Sistema de evaluación

## Convocatoria extraordinaria

- La nota se obtiene como:
- Si  $EEX \geq 4.0$

$$N_{EX} = 0.8 \cdot EEX + 0.2(0.2 \cdot PRA + 0.3 \cdot PRU + 0.45 \cdot EIN + 0.05 \cdot SCO)$$

- Si  $EEX < 4.0$

$$N_{EX} = EEX$$

- Siendo:
  - $EIN$ : Examen intersemestral
  - $EEX$ : Examen extraordinario
  - $PRA$ : Prácticas laboratorio con Matlab (media 4 mejores)
  - $PRU$ : Pruebas cortas clase (media)
  - $SCO$ : Participación activa en clase

*Thank you for your  
attention*

Eugenio Sánchez Úbeda