MAT-042: Probabilidad y Estadística Industrial

Felipe Osorio

http://fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Información

Horario:

Clases: Viernes, bloque 16-19 (19:05-21:45 hrs.), Sala P-219
Taller: Martes, bloque 16-19 (19:05-21:45 hrs.), Sala P-208

Contacto:

E-mail: felipe.osorios@usm.cl.
Web: http://fosorios.mat.utfsm.cl/teaching.html y **AULA**

Evaluación:

Se realizará **3 Certámenes**, **Controles** y **Tareas**.

Ponderaciones:

Sea $\overline{C}$, $\overline{Q}$ y $\overline{T}$ el promedio de certámenes, controles, y tareas, respectivamente. De este modo, la nota de presentación (NP) es dada por:

$$NP = 0.8\,\overline{C} + 0.1\,\overline{Q} + 0.1\,\overline{T}.$$

Criterio de aprobación

Criterio de aprobación:

Aquellos estudiantes que obtengan NP mayor o igual a 55 y **todos** los certámenes sobre 40, aprobarán la asignatura con nota final, $NF = NP$.

Criterio para rendir global:

En caso contrario, y siempre que $NP \geq 45$, los estudiantes podrán rendir el certamen global (CG), en cuyo caso la nota final es calculada como sigue:

$$NF = 0.6 \cdot NP + 0.4 \cdot CG.$$

Reglas adicionales

- Se llevará un control de asistencia.
- Se puede realizar preguntas sobre la materia en cualquier momento.
- Los alumnos deben apagar/silenciar sus teléfonos celulares durante clases.
- Conversaciones sobre asuntos ajenos a la clase no serán tolerados. Otros estudiantes tiene derecho a asistir clases en silencio.
- Al enviar algún e-mail al profesor, identificar el código de la asignatura en el asunto (MAT206).
- E-mail será el canal de comunicación oficial entre el profesor y los estudiantes.

Reglas: sobre las pruebas

- Todas las hojas necesarias para responder las pruebas serán entregadas por el profesor.
- Será permitido el uso de una calculadora científica simple (no del celular).
- Es derecho del estudiante conocer la pauta de corrección la que será publicada en la página web del curso.
- El uso de lápiz grafito es aceptado. Sin embargo, inhabilita al estudiante de pedir recorrección.
- Pedidos de recorrección deben ser argumentados por escrito.
- En modalidad online, Certámenes, Controles y Tareas deben ser enviados en formato **PDF**.[1]
- Cualquier tipo de fraude en prueba (copia, uso de WhatsApp, suplantación, etc.) será llevado a Comisión Universitaria.

[1] En un único archivo, orientado en una dirección legible.

Orientaciones de estudio

- Mantener la frecuencia de estudio de inicio a final del semestre. El ideal es estudiar el contenido luego de **cada** clase.
- Estudiar primeramente el contenido dado en clases, buscando apoyo en las referencias bibliográficas.
- Las referencias son fuentes de ejemplos y ejercicios. Resuelva una buena cantidad de ejercicios. No deje esto para la víspera de la prueba.
- Buscar las referencias bibliográficas al inicio del semestre, dando preferencia a las principales y complementarias.

Programa del curso

- Introducción y conceptos básicos.
- Estadística descriptiva.
- Cálculo de probabilidades.
- Variables aleatorias.
- Inferencia estadística.

Bibliografía

Canavos, G. (1990).
Probabilidad y Estadística, Aplicaciones y Métodos.
McGraw-Hill Latinoamericana.

Meyer, P.L. (1976)
Probabilidad y Aplicaciones Estadísticas.
Fondo Educativo Interamericano.

Newbold, O., Carlson, W.L., Thorne, B. (2008).
Estadística para Administración y Economía (6ta Ed.).
Prentice Hall, Madrid.

Wackerly, D., Mendenhall, W., Scheaffer, R. (2008).
Estadística Matemática con Aplicaciones.
Cengage Learning.

Motivación mediante ejemplos

- ¿Existe competencia en el mercado de AFPs chileno?
- Modelando rentabilidades de acciones en el mercado chileno usando el CAPM[2].
- Ideas sobre el proceso de modelación.
- Algunos conceptos preliminares.

[2] CAPM: Capital asset pricing model

Ideas subyacentes

"Todos los modelos son errados, pero algunos son útiles."

– George Box.

"Aunque puede parecer una paradoja, toda la ciencia exacta está dominada por la idea de aproximación."

– Bertrand Russell.

Principio KISS: "Keep It Simple, Stupid."

– Clarence "Kelly" Johnson.

Ideas subyacentes

"Todos los modelos son errados, pero algunos son útiles."

– George Box.

"Aunque puede parecer una paradoja, toda la ciencia exacta está dominada por la idea de aproximación."

– Bertrand Russell.

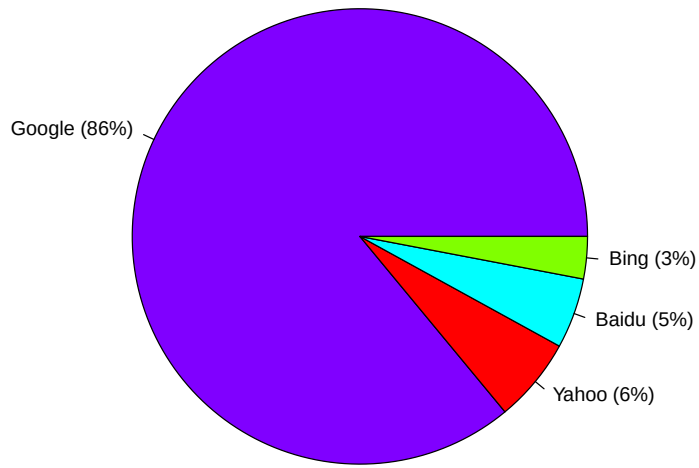
Principio KISS: "Keep It Short and Simple."

– Clarence "Kelly" Johnson.

El éxito de Google: Aplicar el principio KISS[3]

Evolución de Yahoo vs. Google:

Cuota de mercado de los motores de búsqueda:



[3] En estadística este se conoce como Principio de Parsimonia.

Esto NO es una crítica al sistema de AFP...

Administradoras de Fondos de Pensiones (AFP) de Chile

Aplicación:

El sistema de AFP (o de capitalización individual) chileno está en vigor desde 1980.

Ahorros de los contribuyentes son administrados en un sistema de multifondos.

Existe 5 tipos de fondos (A, B, C, D y E) divididos por la proporción del portfólio que es invertido en títulos de renta variable.

El fondo A tiene la mayor proporción de inversión en renta variable, la que disminuye progresivamente para los fondos B, C, D y E.

Conjunto de datos:

Rentabilidades mensuales de AFPs: Cuprum, Habitat, PlanVital y ProVida en el periodo de agosto/2005 a abril/2020.

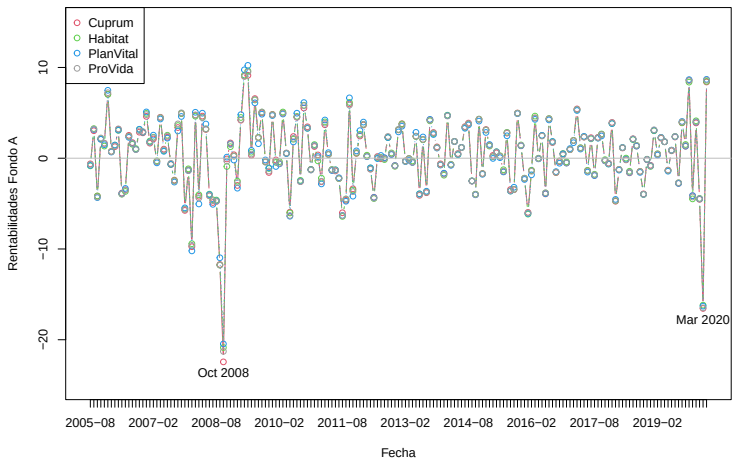
Datos fueron obtenidos desde el sitio web de la superintendencia de pensiones (www.spensiones.cl)

Conjunto de datos con 177 observaciones y 4 variables (para cada uno de los fondos).

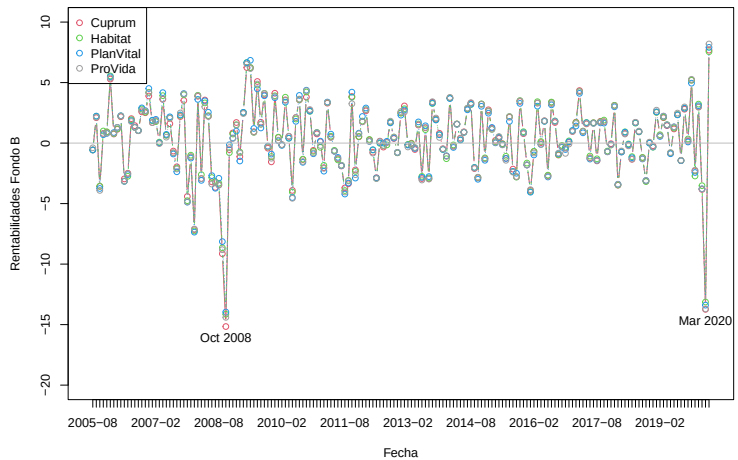
Varias observaciones son identificadas como outliers.

QQ-plot de distancias transformadas revelan la presencia de colas pesadas.

Rentabilidades de AFPs chilenas

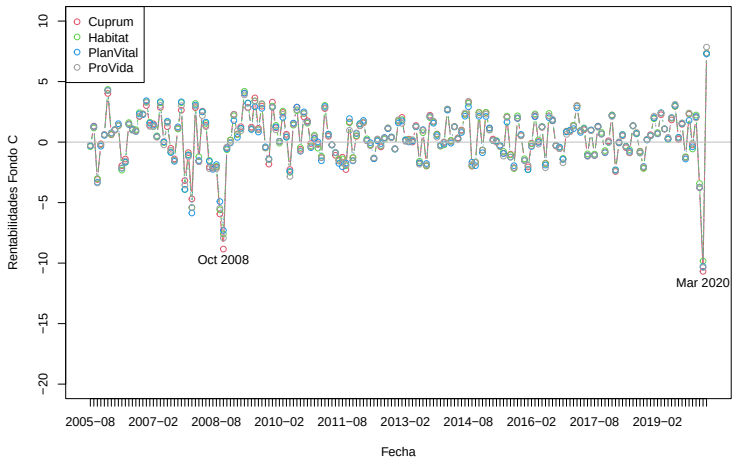


(a) Fondo A



(b) Fondo B

Rentabilidades de AFPs chilenas



(a) Fondo C

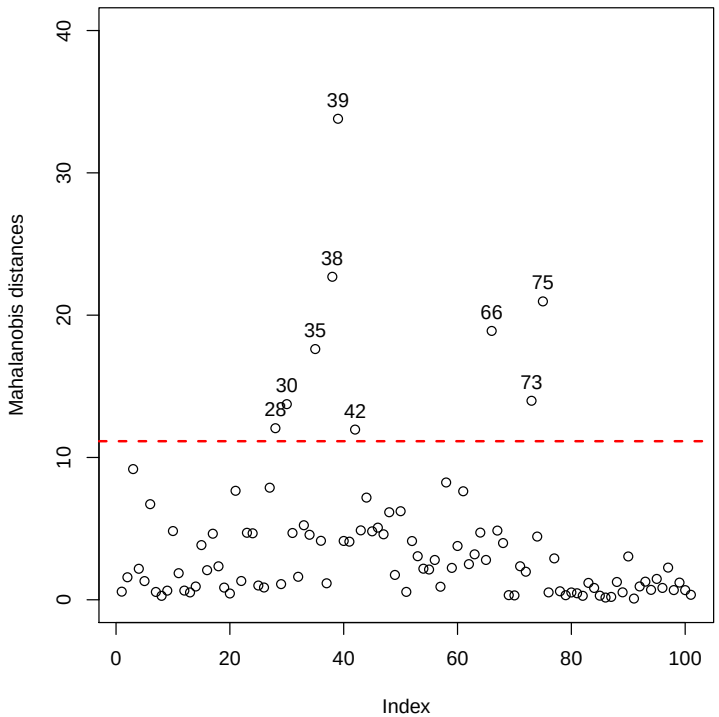
(b) Fondo E

Identificando observaciones atípicas

En mercados emergentes como el chileno suele ocurrir periodos con alta volatilidad.

Existe una bateria de procedimientos para detectar observaciones que presentan un comportamiento es aberrante/atípico.

Este tipo de observaciones puede tener un efecto nefasto sobre la inferencia estadística.

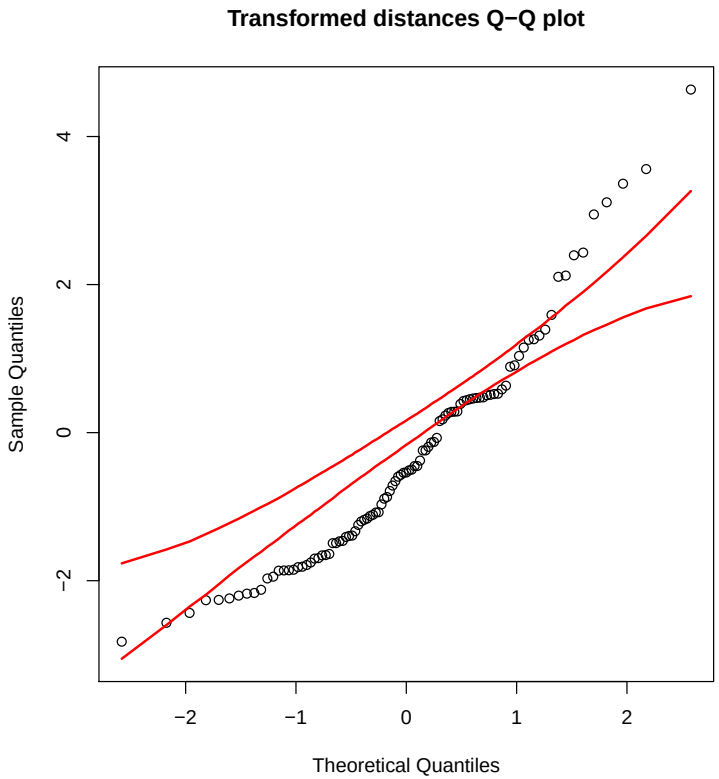


Evaluando los supuestos distribucionales

El supuesto de normalidad es habitual en este tipo de problemas.

Es decir, suponga $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ una muestra aleatoria desde $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Usando test de hipótesis y técnicas gráficas se concluye que el supuesto de normalidad no es soportado por los datos.



Análisis multivariado usando la distribución t de Student

Características del problema:

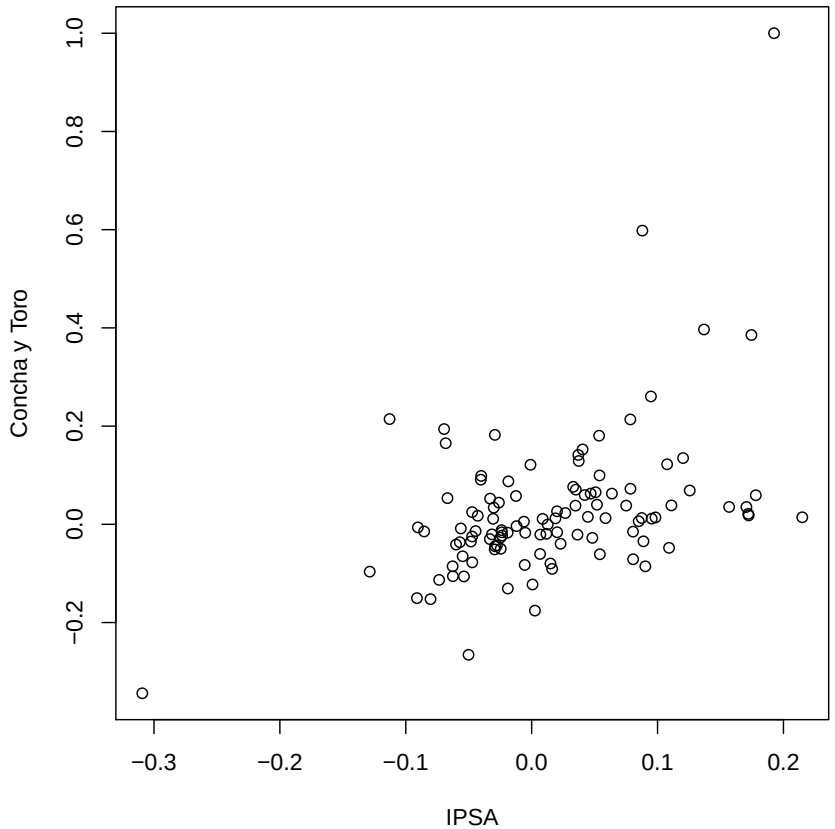
- AFPs invierten esencialmente en la misma cartera de inversiones.
- Mercados emergentes suelen presentar alta volatilidad.
- Los datos son bien modelados usando distribuciones con colas pesadas.

Conclusiones:

- Aparentemente, no existe competencia en el mercado de AFP.
- Cálculo óptimo de los porcentajes de inversión en los distintos fondos.
- Evaluar la igualdad entre razones de Sharpe.

Datos de Concha y Toro (Osorio y Galea, 2006)[4]

Rentabilidades mensuales de Concha y Toro vs. IPSA, ajustados por bonos de interés del Banco Central entre marzo/1990 a abril/1999.



[4] Statistical Papers **47**, 31-38

Datos de Concha y Toro

Modelo CAPM (Valoración de Activos de Capital), Sharpe (1964)[5]

$$\mathrm{E}(r) = r_f + \beta(\mathrm{E}(r_m) - r_f),$$

usando datos observados, podemos escribir

$$R_t = \alpha + \beta \times IPSA_t + \epsilon, \qquad t = 1, \ldots, T.$$

Características del problema:

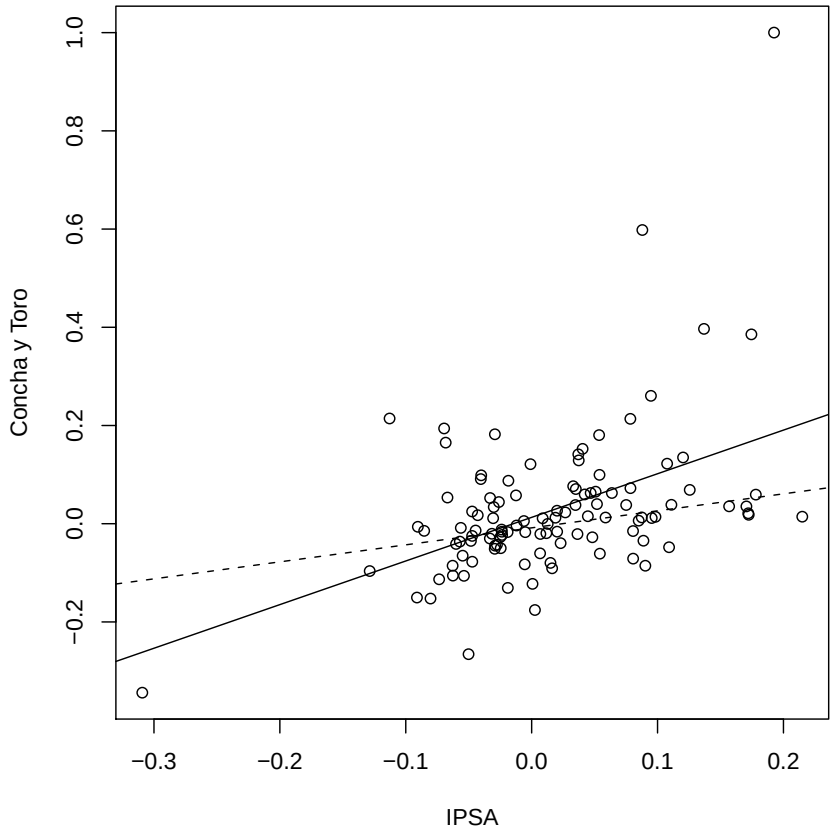
- Relación lineal entre las variables.
- Posibles periodos de alta volatilidad.

Hipótesis de interés:

- $H_0 : \beta > 1$ (Amante del riesgo).
- $H_0 : \beta = 1$ (Neutral al riesgo).
- $H_0 : \beta < 1$ (Averso al riesgo).

[5] Journal of Finance **19**, 425-442

Datos de Concha y Toro



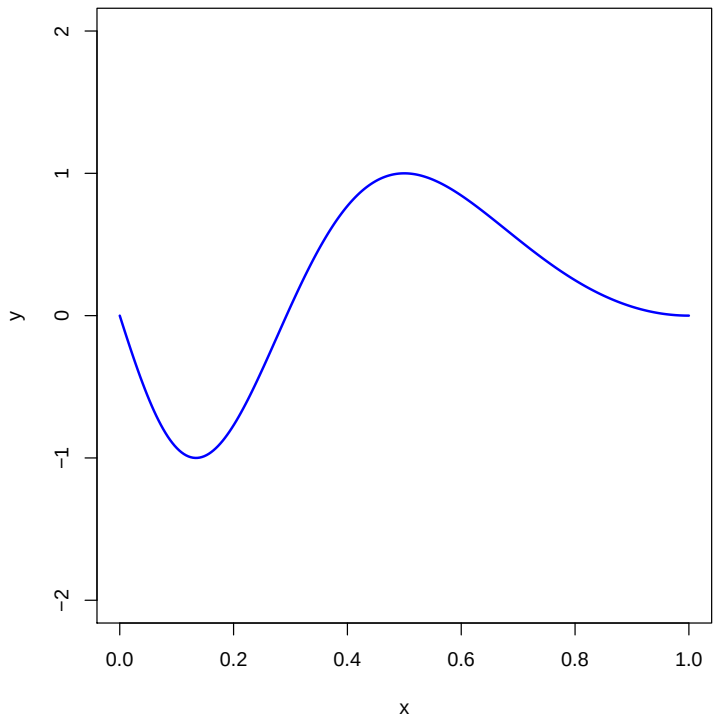
Ajuste usando errores normales (—) y Cauchy (– –). ($\widehat{\beta} = 0.89$ y $\widehat{\beta} = 0.35$, respectivamente).

El problema del modelado

Considere la función

$$Y = \text{sen}\{2\pi(1-x)^2\},$$

cuyo gráfico es dado por:

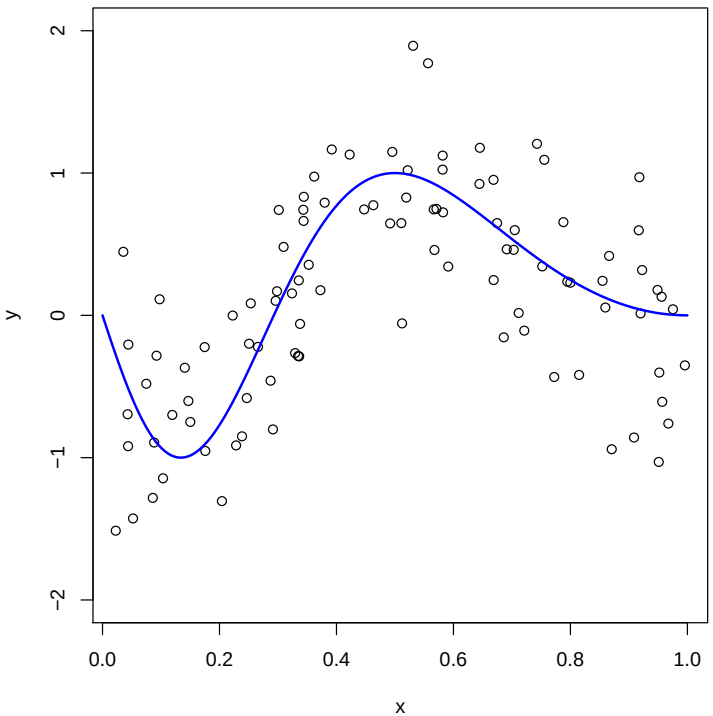


El problema del modelado

Suponga que "generamos" datos, usando

$$Y_i = \text{sen}\{2\pi(1 - x_i)^2\} + \sigma\epsilon_i, \qquad i = 1, \ldots, 100,$$

donde $x_i \sim \mathcal{U}(0, 1)$, $\epsilon_i \sim \mathcal{N}(0, 1)$ y $\sigma = 1/2$,

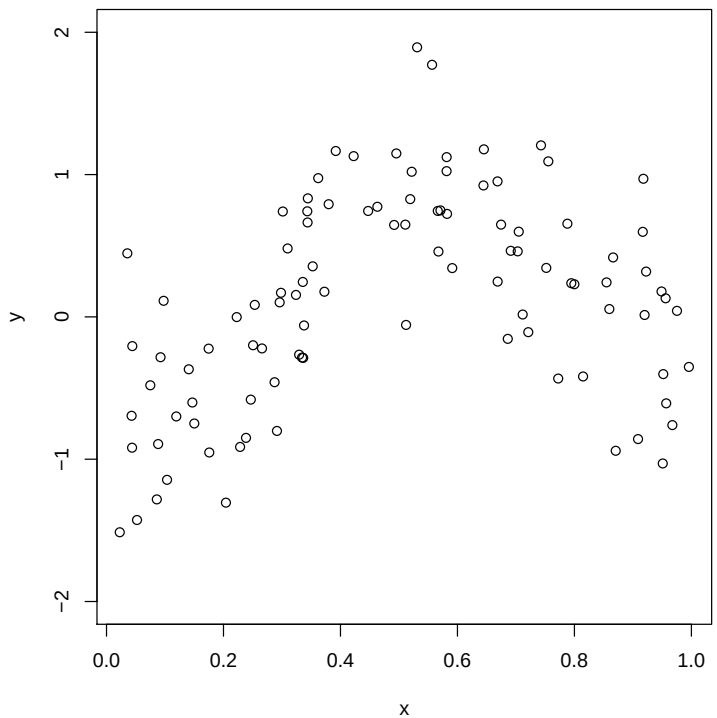


El problema del modelado

Lamentablemente, en la práctica **sólo** disponemos de los datos observados:

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_{100}, Y_{100}),$$

el primer paso es hacer un análisis exploratorio:

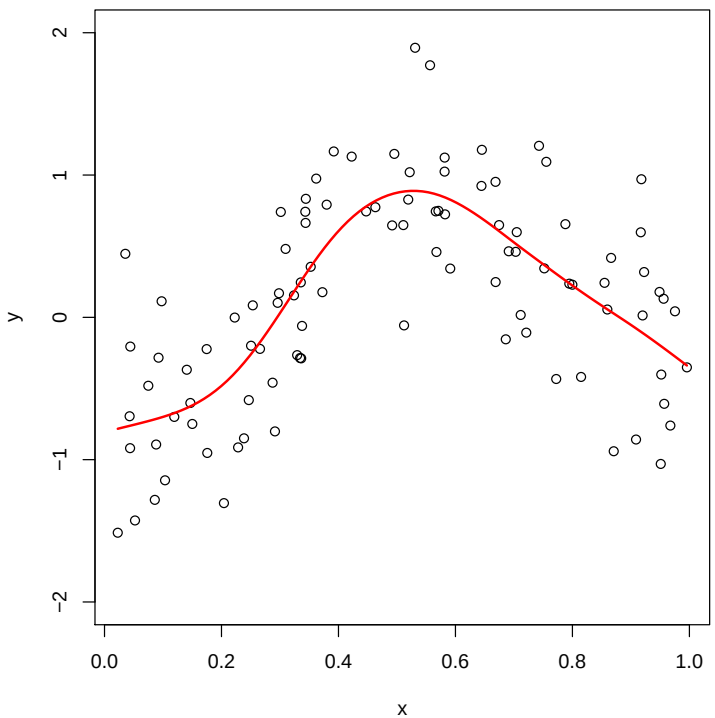


El problema del modelado

El analista propone el modelo:

$$Y_i = g(x_i) + \epsilon_i, \qquad i = 1, \ldots, 100,$$

y su objetivo es "estimar" la función $g(\cdot)$ desde los datos, obteniendo

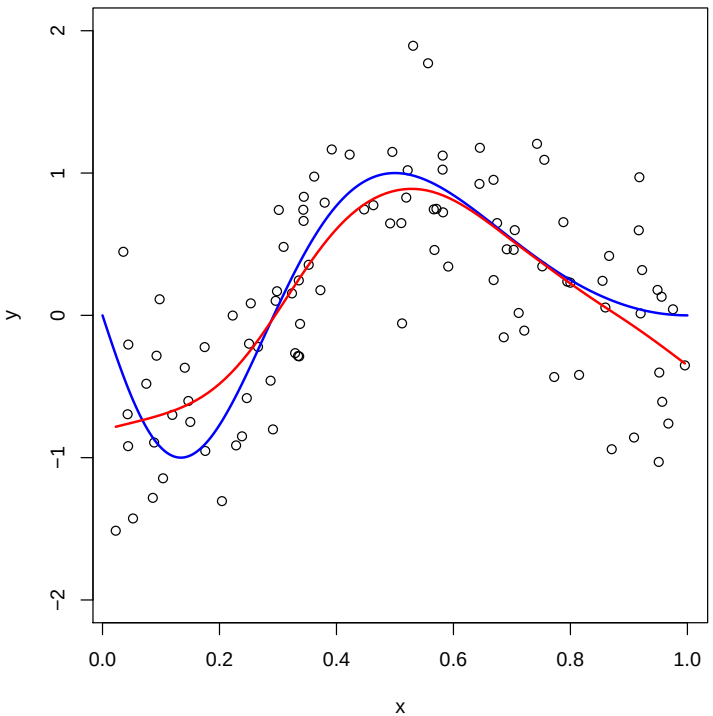


El problema del modelado

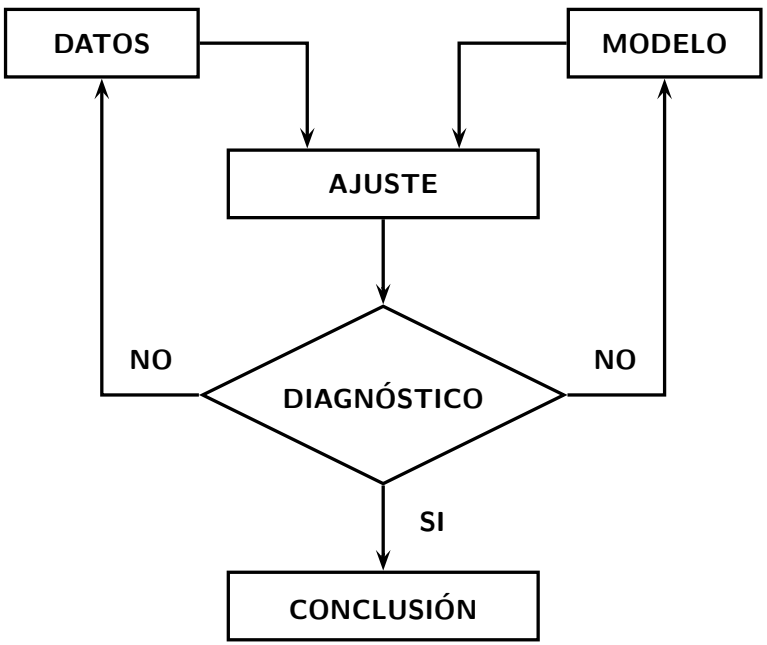
En Estadística se estudia teóricamente, la "bondad del modelo" comparando

$$\widehat{Y} = \widehat{g}(x), \qquad \text{v.s.} \qquad Y = \text{sen}\{2\pi(1-x)^2\},$$

esto es, el modelo ajustado v.s. el modelo subyacente (verdadero).



Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

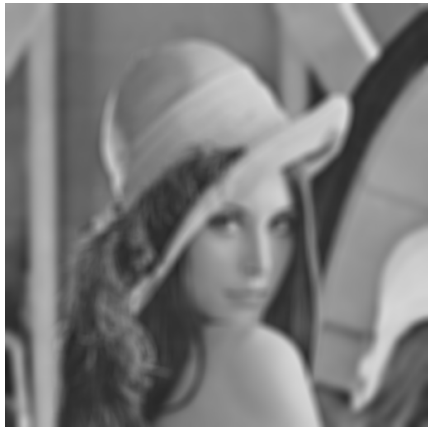
Lenna y algunas distorsiones de Lenna

(a) Original image

(b) Mean shift

(c) Salt-pepper noise

(d) Speckle noise



(e) Blurring

(f) Compression

Similaridad entre imágenes

- Existen diversos enfoques para estudiar la similaridad entre dos señales, imágenes o (en general) procesos.
- El objetivo de la evaluación de la calidad de una imagen busca representar la percepción de la calidad del ojo humano.
- Se ha diseñado índices para estudiar el desempeño de algoritmos para problemas como: compresión o restauración de imágenes, entre otros. Algoritmos de referencia completa requieren de imágenes distorcionadas **y** de referencia.
- Se desea un coeficiente apropiado que combine la luminosidad, contraste y estructura (correlación) entre las imágenes. Este tipo de coeficientes son llamados índice de similaridad estructural (SSIM).

Structural Similarity Index (SSIM)

Definición (Wang et al., 2004):[6]

Sean $\boldsymbol{x}, \boldsymbol{y}$ dos imágenes. El índice SSIM es definido como

$$\mathrm{SSIM}(\boldsymbol{x}, \boldsymbol{y}) = l(\boldsymbol{x}, \boldsymbol{y})^{\alpha} \cdot c(\boldsymbol{x}, \boldsymbol{y})^{\beta} \cdot s(\boldsymbol{x}, \boldsymbol{y})^{\gamma},$$

donde α, β y γ son parámetros no negativos,

$$l(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\,\overline{x}\,\overline{y} + c_1}{\overline{x}^2 + \overline{y}^2 + c_1}, \qquad c(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\,s_x\,s_y + c_2}{s_x^2 + s_y^2 + c_2},$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{s_{xy} + c_3}{s_x\,s_y + c_3},$$

$\overline{x}$, $\overline{y}$, s_x^2, s_y^2 y s_{xy} representan los promedios muestrales, varianzas y covarianza de $\boldsymbol{x}$ y $\boldsymbol{y}$.

Las constantes c_1, c_2 y c_3 garantizan la estabilidad cuando denominadores son cercanos a cero.

[6] IEEE Transactions on Image Processing **13**, 600-612.

¿Cómo lucen los datos de Lenna?[7]

Lenna (original):

$$\begin{pmatrix} 153 & 153 & 153 & 152 & 153 & \dots \\ 153 & 153 & 153 & 152 & 153 & \dots \\ 153 & 153 & 153 & 152 & 153 & \dots \\ 153 & 153 & 153 & 152 & 153 & \dots \\ 153 & 153 & 153 & 152 & 153 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Lenna (sal y pimienta 10% contaminación):

$$\begin{pmatrix} 153 & 153 & 153 & 152 & 153 & \dots \\ 153 & 153 & 30 & 152 & 153 & \dots \\ 153 & 153 & 153 & 152 & 153 & \dots \\ 153 & 153 & 62 & 152 & 153 & \dots \\ 66 & 153 & 153 & 152 & 153 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

[7] Imágenes $512 \times 512 = 262\,144$ observaciones.

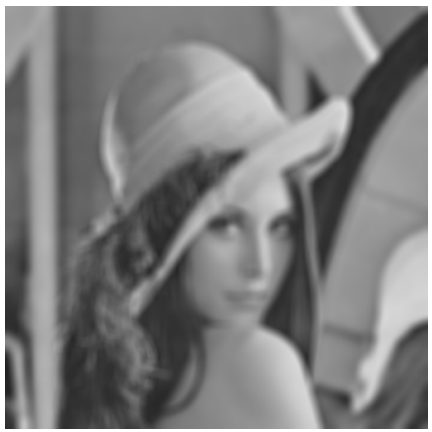
Lenna y algunas distorsiones de Lenna[8]

(a) Original image

(b) Mean shift

(c) Salt-pepper noise

(d) Speckle noise



(e) Blurring

(f) Compression

[8] SSIM: (a) 1.000, (b) 0.989, (c) 0.649, (d) 0.441, (e) 0.346 y (f) 0.288.

Resumiendo Lena

Para los datos de Lena (original) podemos calcular, por ejemplo:

$$\begin{aligned}\overline{x} &= \frac{1}{262144}(153 + 153 + \cdots + 69 + 76 + 77 + 89 + 89) \\ &= 100.0519,\end{aligned}$$

adicionalmente, el rango en que fluctuan los datos de Lena es $[0, 255]$. Es decir,[9]

$$\min\{x_1, x_2, \ldots, x_{262144}\} = 0, \qquad \min\{x_1, x_2, \ldots, x_{262144}\} = 255.$$

¿Cónoce otras medidas de resumen?

[9] En esta escala de grises, 0 indica el negro, mientras que 255 el blanco.