MAT-269: Análisis Estadístico Multivariado

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Información

Horario:

Clases: Lunes y Miércoles, bloque 7-8 (12:15-13:25 hrs.), salas M101 y M104.

Contacto:

E-mail: felipe.osorios@usm.cl.

Material de clases:

Página del curso (GitHub): https://github.com/faosorios/Curso-Multivariado
Página personal: http://fosorios.mat.utfsm.cl/teaching.html#MAT269
El material también estará disponible en AULA.

Evaluación:

Se realizará **3 Certámenes**.

Criterio de aprobación

Criterio de aprobación:

Sea NP el promedio de los Certámenes. 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 $NP \geq 45$,[1] los estudiantes podrán rendir el certamen global (CG), en cuyo caso la nota final (NF) es calculada como sigue:

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

[1] Si $NP < 45$ usted ha reprobado la asignatura.

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 (MAT269).
- E-mail será el canal de comunicación oficial entre el profesor y los estudiantes.

Reglas: sobre los certámenes

- Es derecho del estudiante conocer la pauta de corrección la que será publicada en la página web del curso.
- Use principalmente lapiz pasta (no utilice lapiz rojo).
- Pedidos de recorrección deben ser argumentados por escrito.
- En modalidad online, Certámenes deben ser enviados en formato **PDF**.[2]
- Cualquier tipo de fraude en prueba (copia, uso de WhatsApp, suplantación, etc.) será llevado a Comisión Universitaria.

[2] 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.

Prerrequisitos

- El requisito formal es MAT-266: Análisis de Regresión.
- Adicionalmente *[el profesor tiene la muy mala costumbre de..]* usaremos algunas ideas desde MAT-206: Inferencia Estadística.
- Se asume un conocimiento básico de los siguientes aspectos:
  - Vectores aleatorios.
  - Distribución normal multivariada.
  - Manipulación de matrices y vectores aleatorios.

Programa del curso[3]

- Inferencia en análisis multivariado.
  - Estimación y test de hipótesis para una muestra aleatoria desde $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Técnicas multivariadas.
  - Regresión multivariada y GMANOVA.
  - Análisis de componentes principales.
  - Análisis factorial.
  - Métodos de clasificación y agrupamiento.
- Tópicos adicionales.*

[3] Este es un curso fundamental donde exploramos métodos para abordar la inferencia estadística en análisis multivariado, no es un curso enfocado en el análisis de datos.

Referencias bibliográficas


Anderson, T.W. (2003).
An Introduction to Multivariate Statistical Analysis (3rd Ed.).
Wiley, New York.


Härdle, W.K., Simar, L. (2012).
Applied Multivariate Statistical Analysis (3rd Ed.).
Springer, New York.


Seber, G.A.F. (2004).
Multivariate Observations.
Wiley, New York.

Motivación mediante ejemplos

- ¿Existe competencia en el mercado de AFPs chileno?
- El desempeño de los estudiantes chilenos en el SIMCE.
- Un problema de clasificación clásico, o por qué nos presta dinero el banco.
- Recordatorio: el esquema de modelación.

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.

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.

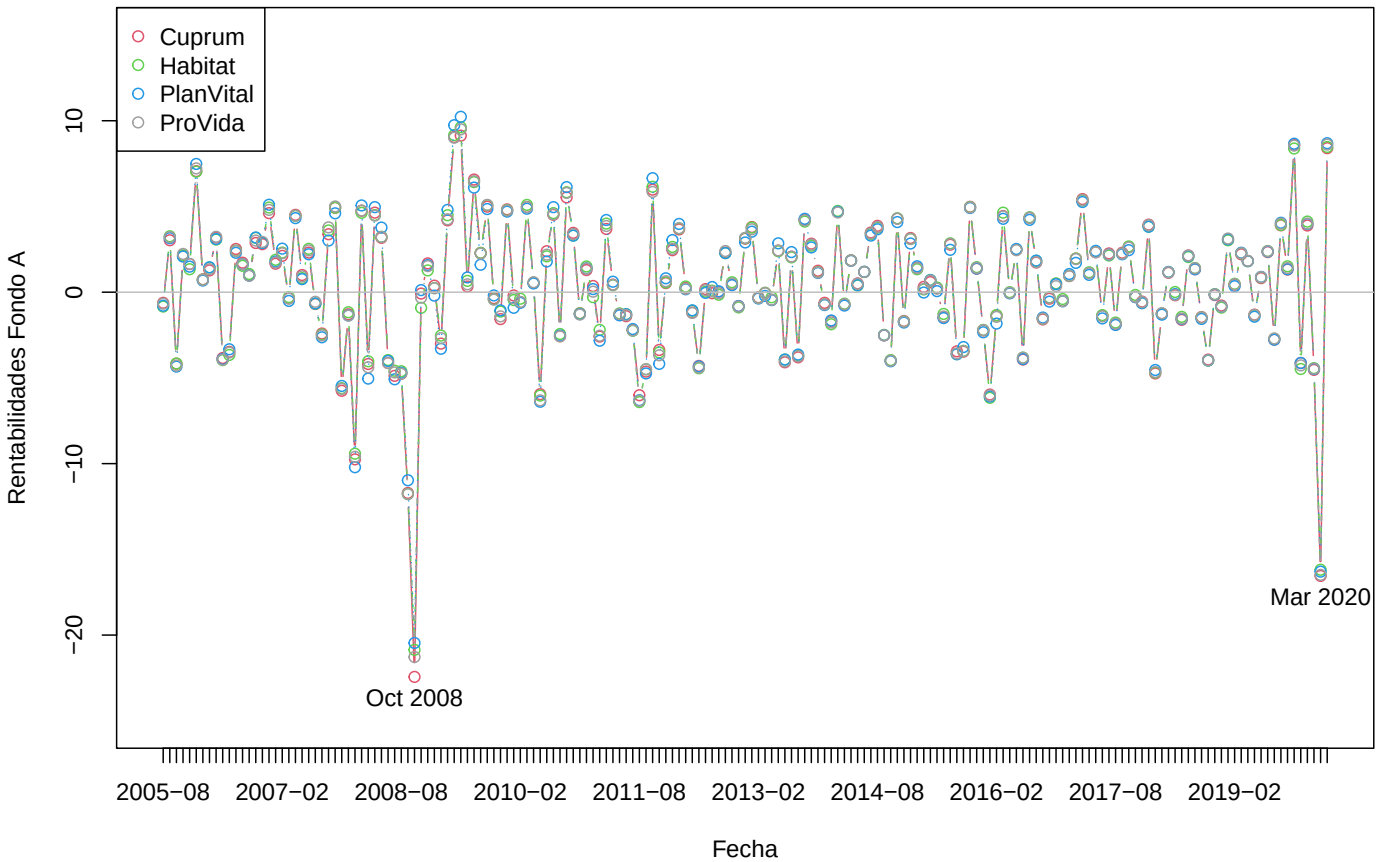
Datos fueron obtenidos desde el sitio web de la superintendencia de pensiones (www.spensiones.cl)

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

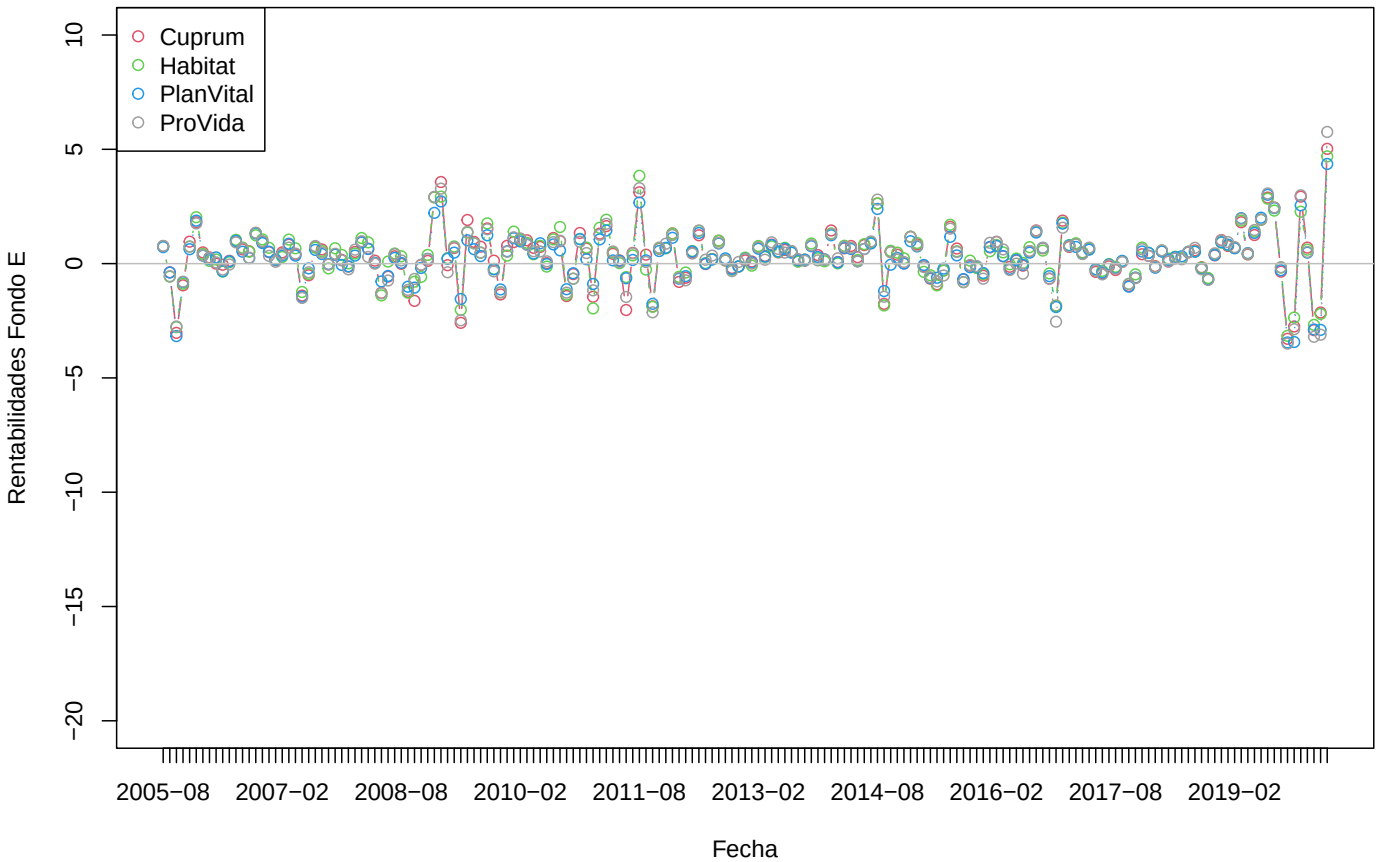
Varias observaciones son identificadas como outliers.

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

Rentabilidades de AFPs chilenas



Rentabilidades de AFPs chilenas

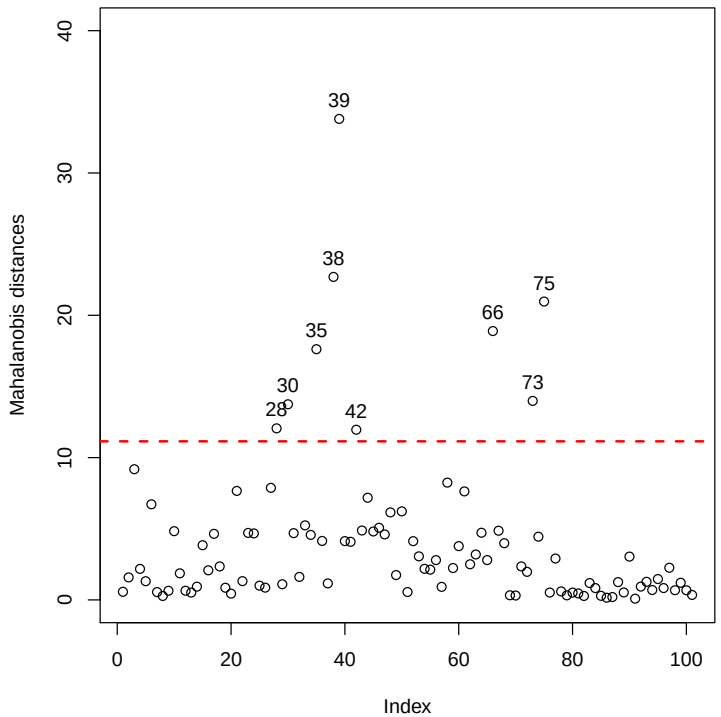


Identificando observaciones atípicas

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.[4]



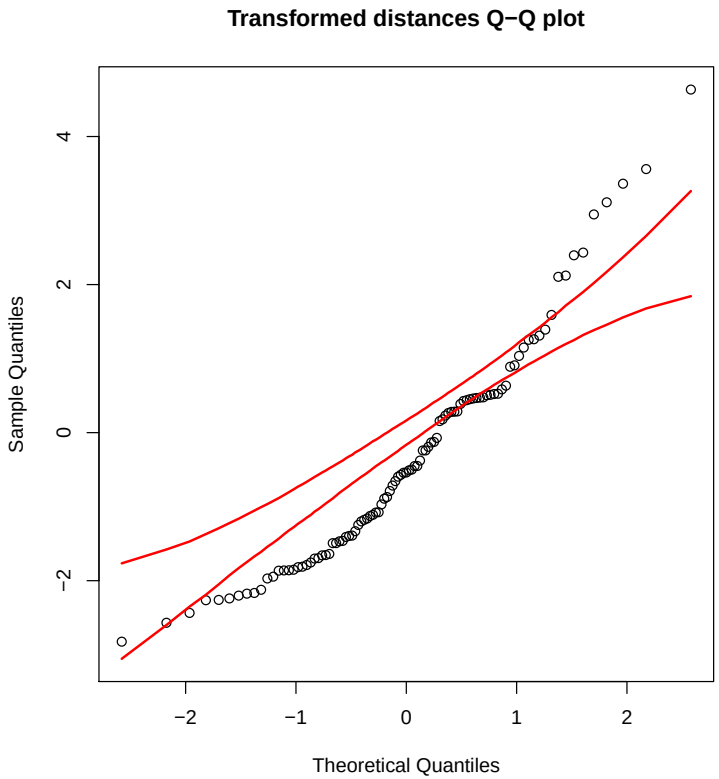
[4] Es decir, Ud. puede llegar a conclusiones erróneas!

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 del SIMCE

Desde 1988 el SIMCE evalua los resultados de aprendizaje de los estudiantes del sistema de educación chileno.

Objetivos:

- Describir el comportamiento del aprendizaje de los estudiantes.
- Determinar si existe diferencias significativas entre el tipo de dependencia (municipal, subvencionado, particular).

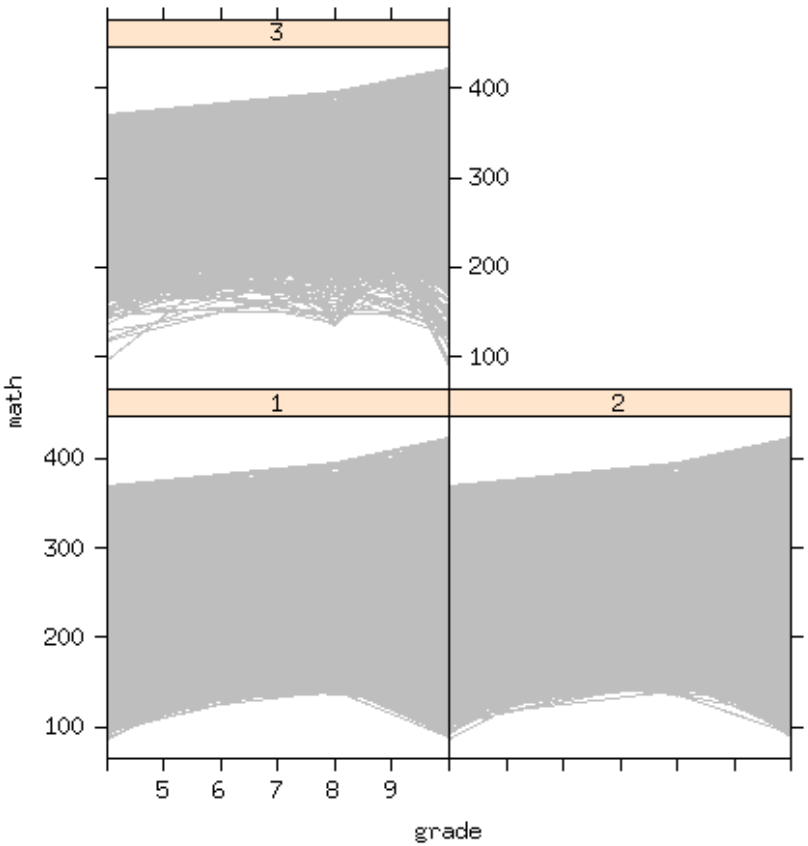
Características del problema:

- Mediciones de un mismo individuo (estudiante) a través del tiempo (4º y 8º básico, 2º medio).[5]
- Datos disponibles para los años 2007, 2011 y 2013, pruebas de Lenguaje y Matemáticas.

[5] Conocido como: datos con estructura longitudinal.

Datos del SIMCE

Perfiles individuales de los puntajes del SIMCE en matemáticas, organizados por tipo de dependencia.



Estimación robusta en modelos de curvas de crecimiento

Características del problema:

- Aproximadamente 132K estudiantes para ser analizados (base de datos de mediano porte).
- Crecimiento lineal (cuadrático?) a través del tiempo.
- Igual número de mediciones por individuo (datos balanceados).

Alternativas para análisis:

- Modelos con efectos-mixtos.
- Modelos multi-nivel.
- Modelo de curvas de crecimiento (GMANOVA).

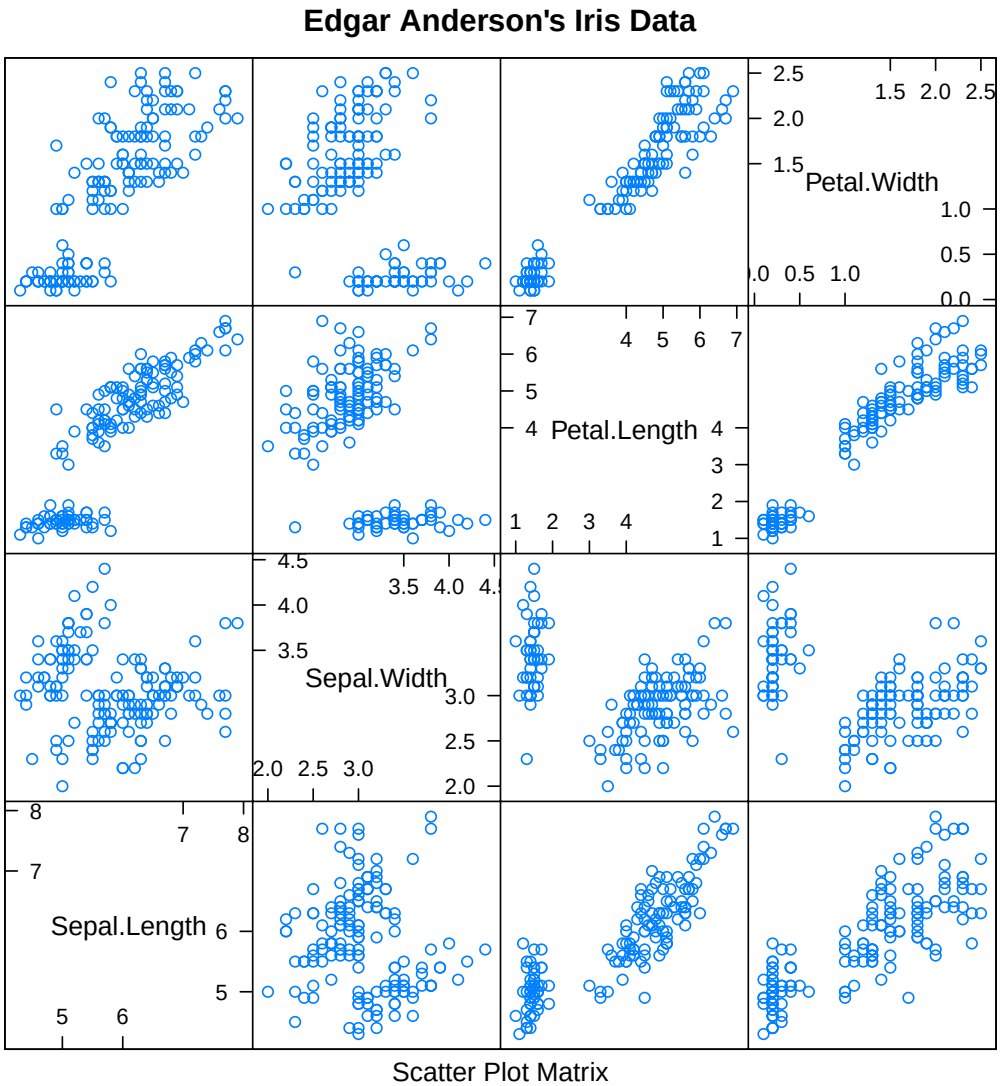
Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Iris setosa

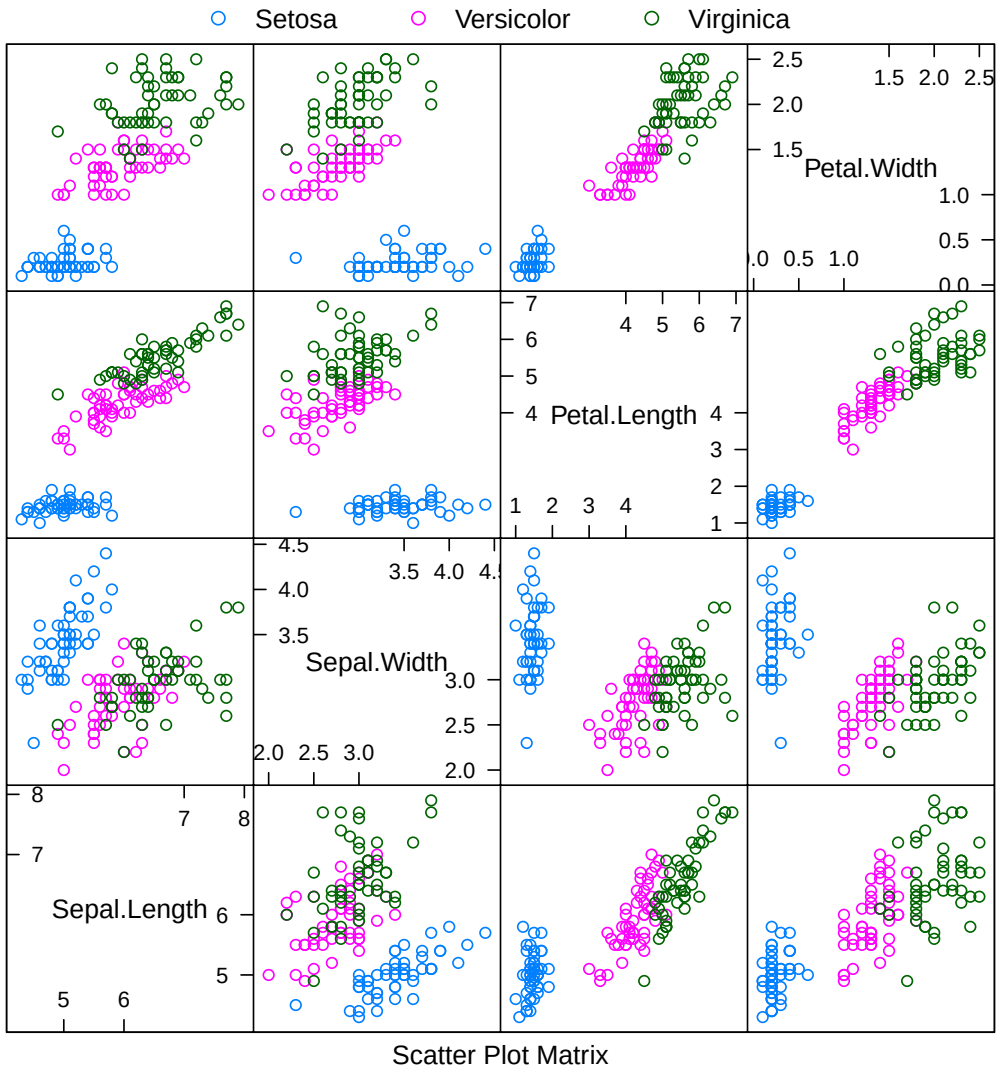
Iris versicolor

Iris virginica

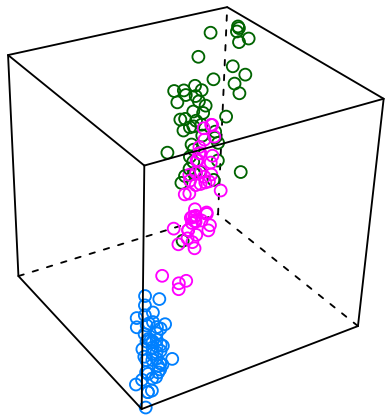
Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

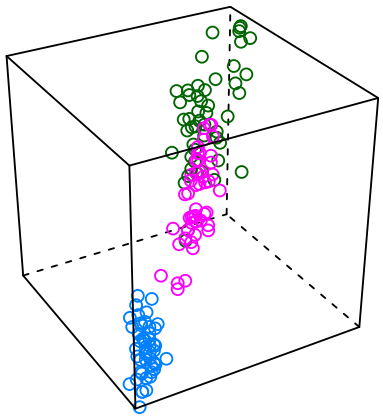


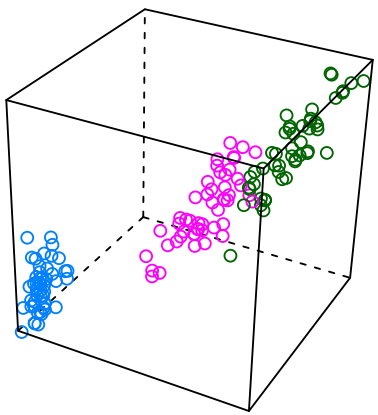
Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

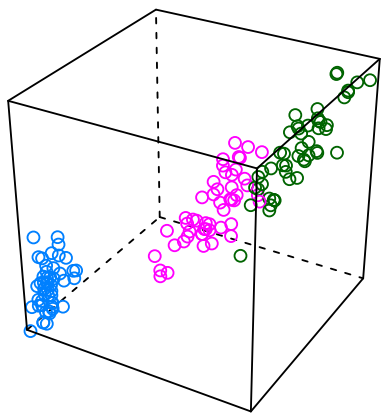


Datos de Flores Iris (Anderson, 1935; Fisher, 1936)









Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

Objetivo:

- Obtener una función que permita discriminar entre especies.
- Usando las medidas de una flor, clasificarla apropiadamente.

Características del problema:

- El análisis exploratorio revela una separación evidente en 2 grupos.
- Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
  - Análisis discriminante,
  - Técnicas de clasificación (Reconocimiento de patrones),
  - Aprendizaje de máquina (Máquinas de soporte vectorial, Data mining).

Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

Objetivo:

- Obtener una función que permita discriminar entre especies.
- Usando las medidas de una flor, clasificarla apropiadamente.

Características del problema:

- El análisis exploratorio revela una separación evidente en 2 grupos.
- Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
  - Análisis discriminante,
  - Técnicas de clasificación (Reconocimiento de patrones),
  - Aprendizaje de máquina (Máquinas de soporte vectorial, Data mining).

Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

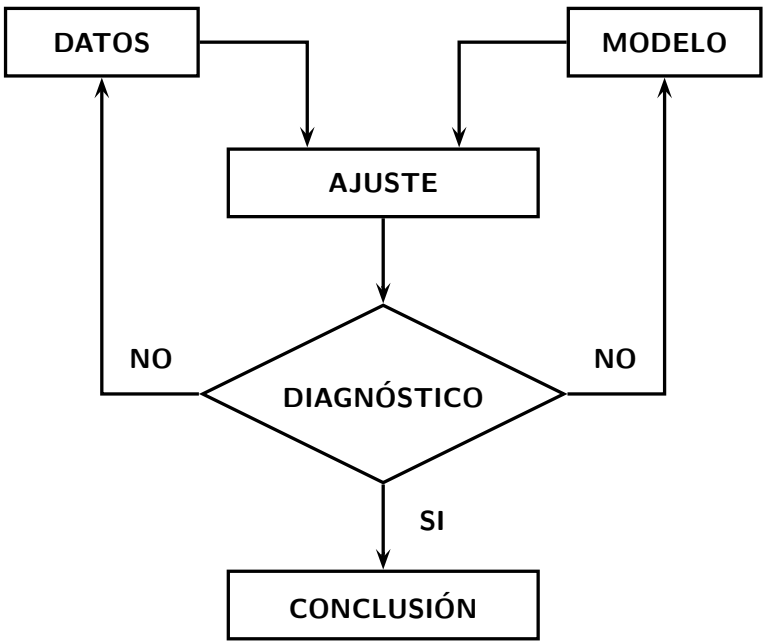
Objetivo:

- Obtener una función que permita discriminar entre especies.
- Usando las medidas de una flor, clasificarla apropiadamente.

Características del problema:

- El análisis exploratorio revela una separación evidente en 2 grupos.
- Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
  - Análisis discriminante,
  - Técnicas de clasificación (Reconocimiento de patrones),
  - Aprendizaje de máquina (Máquinas de soporte vectorial, Data mining).

Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

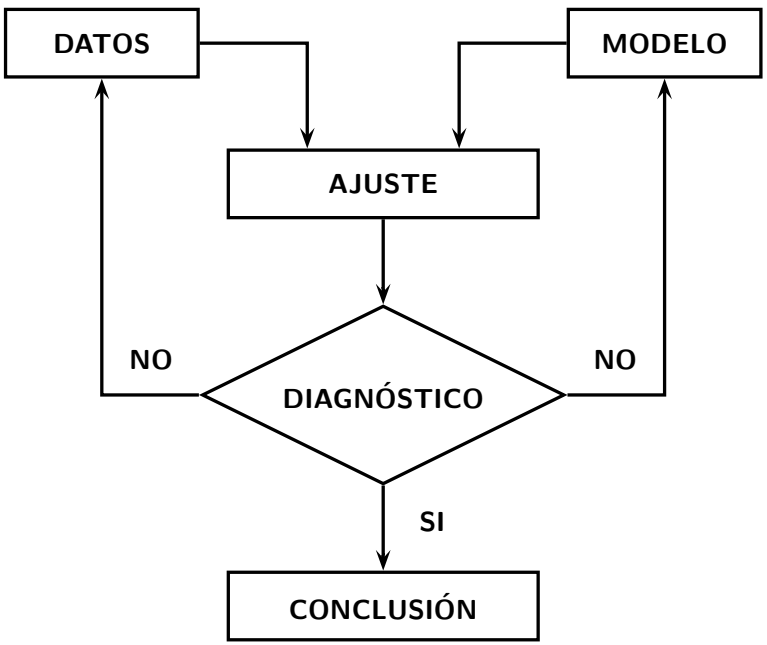
Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

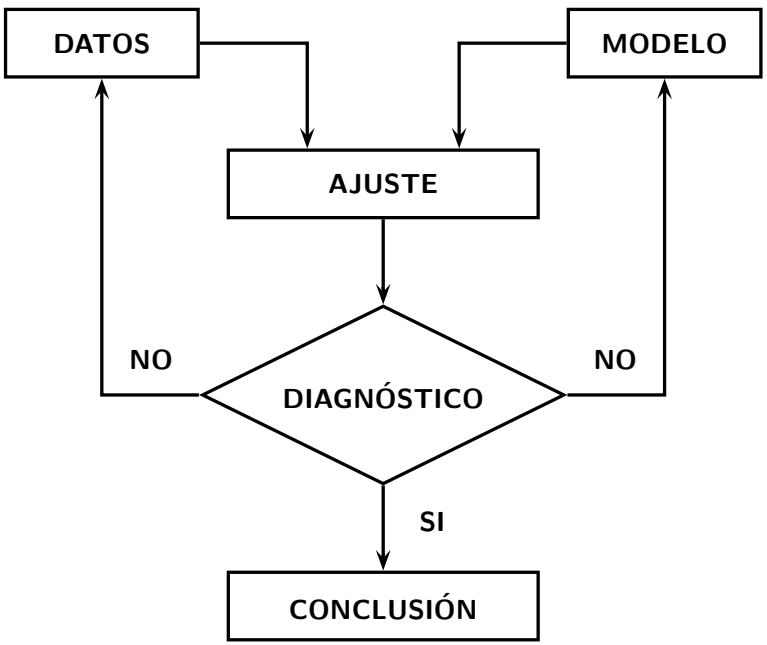
Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

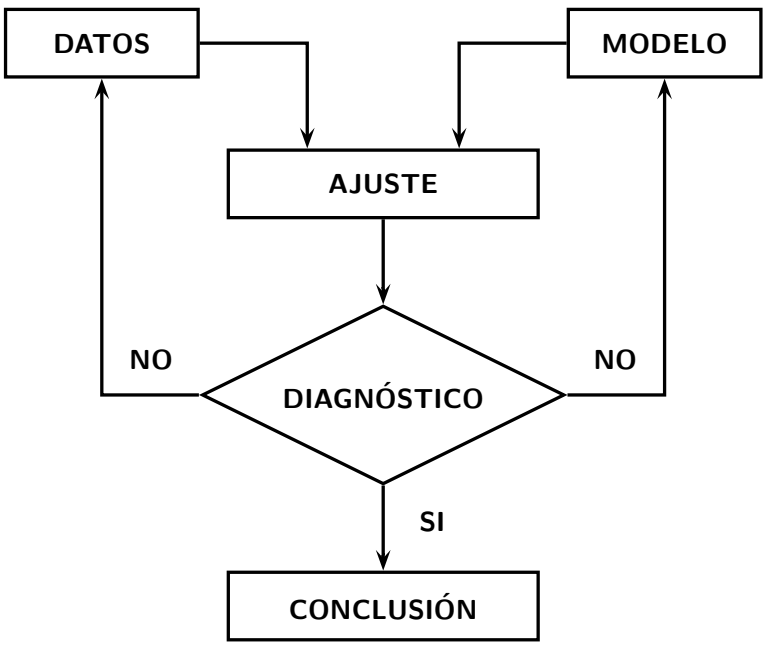
Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

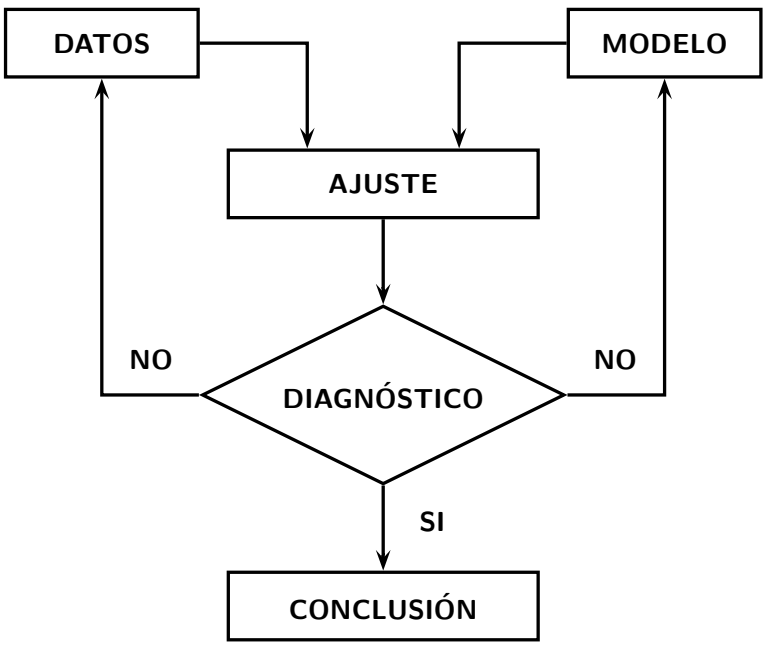
Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

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!