MAT-266: Análisis de Regresión

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Información

Horario:

Clases: Lunes y Miércoles, bloque 3-4 (09:35-10:45 hrs.), salas P111 y P212.

Contacto:

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

Material de clases:

Página del curso (GitHub): `https://github.com/faosorios/Curso-Regresion`
Página personal: `http://fosorios.mat.utfsm.cl/teaching.html#MAT266`

Evaluación:

Se realizará **3 Certámenes** (05-Abril, 10-Mayo, 19-Junio).

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.
- Alumnos que lleguen tarde o se retiren deben hacerlo en silencio.
- Al enviar algún e-mail al profesor, identificar el código de la asignatura en el asunto (MAT266).
- E-mail será el canal de comunicación oficial entre el profesor y los estudiantes.

Reglas: sobre los certámenes

- Todas las hojas necesarias para responder los certámenes 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.
- 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-041: Probabilidad y Estadística.[3]
- Adicionalmente usaremos algunas ideas desde MAT-206: Inferencia Estadística.
- Se asume un conocimiento básico de los siguientes aspectos:
  - Variables aleatorias.
  - Convergencia de variables aleatorias.
  - Manipulación de matrices y vectores aleatorios.

[3] O sus equivalentes MAT-031 o MAT-042.

Programa del curso[4]

1. Preliminares.
2. Inferencia en el modelo de regresión lineal.
3. Análisis de los supuestos del modelo.
4. Alternativas a mínimos cuadrados.
5. Identificación del mejor conjunto de regresores.
6. Tópicos adicionales.

[4] Este es un curso fundamental donde exploramos métodos para abordar el proceso de inferencia en modelos de regresión, **no** es un curso enfocado en el análisis de datos.

Bibliografía

Hocking, R. (2013).
Methods and Applications of Linear Models: Regression and the analysis of variance, 3rd Edition.
Wiley, New York.

Seber, G.A.F., Lee, A.J. (2007).
Linear Regression Analysis, 2nd Edition.
Wiley, New York.

Weisberg, S. (2013).
Applied Linear Regression, 4th Edition.
Wiley, New York.

Objetivo del análisis de regresión

Estudiar una variable de respuesta, y [asuminda continua] como función de algunas variables explicativas o regresores, $x_1, x_2, \dots$ [pueden ser discretas y/o continuas].

$$\textbf{Entrada} \quad \begin{array}{c}\rightarrow\\ \rightarrow\\ \rightarrow\end{array} \quad \boxed{\textbf{Proceso}} \quad \rightarrow \quad \textbf{Salida}$$

En ocasiones la relación funcional es conocida salvo algunos coeficientes (parámetros).

Es decir, la relación es gobernada por un proceso físico o por leyes bien aceptadas

$$Y \approx f(x_1, \dots, x_p; \boldsymbol{\theta}),$$

en cuyo caso, el interés recae en estimar el vector de parámetros $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$.

Modelamiento estadístico

Asumiremos variables aleatorias independientes $Y_1, \dots, Y_n$, tal que

$$Y_i = \mu_i + \epsilon_i, \qquad \mathsf{E}(\epsilon_i) = 0, \qquad i = 1, \dots, n,$$

esto es,

$$\text{respuesta} = \text{parte sistemática} + \text{error aleatorio}$$

Idea:

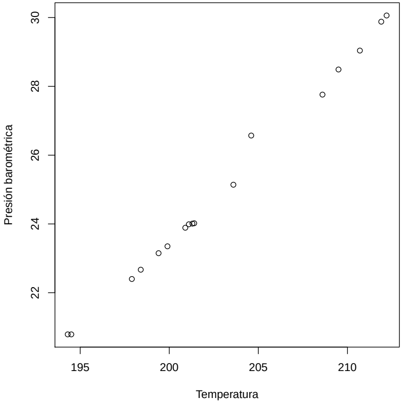
Se desea **"estructurar"** la función de medias como

$$\mu_i = \mu_i(\boldsymbol{\beta}), \qquad i = 1, \dots, n,$$

con $\boldsymbol{\beta} \in \mathbb{R}^p$ y $n \gg p$.

Datos de Forbes (1857)[5]

Presión barométrica en pulgadas de mercurio y temperatura de ebullición del agua en grados Fahrenheit para 17 diferentes altitudes.



[5] Transactions of the Royal Society of Edinburgh **21**, 235-243.

Datos de Forbes

Para describir la relación entre la temperatura y la media de la presión barométrica, podemos considerar

$$\mu = \beta_0 + \beta_1 x,$$

note que

$$\mu = \boldsymbol{x}^\top \boldsymbol{\beta}, \qquad \boldsymbol{x} = (1, x)^\top, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1)^\top,$$

y $\boldsymbol{x}^\top \boldsymbol{\beta}$ se denomina predictor lineal.

El conjunto de datos consiste del vector de respuestas.

$$\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top,$$

y una matriz de diseño $n \times 2$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}.$$

Convención:

Todos los vectores siempre serán columna.

Datos de Forbes[7]

Considere el modelo

$$\text{Presión}_i = \beta_0 + \beta_1 \,\text{Temperatura}_i + \epsilon_i,$$

para $i = 1, \dots, 17$.

En nuestro caso (usando función `lm` de **R**[6]), obtuvimos:

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^\top = (-81.0637, 0.5229)^\top$$

$$s^2 = \frac{1}{17-2} \sum_{i=1}^{n} (\text{Presión}_i - \widehat{\beta}_0 - \widehat{\beta}_1 \,\text{Temperatura}_i)^2$$

$$= 0.0542.$$

Además, $R^2 = \mathsf{cor}^2(\text{Presión}, \widehat{\text{Presión}}) = 0.9944$, donde

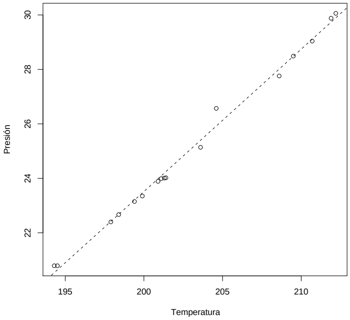
$$\widehat{\text{Presión}}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \,\text{Temperatura}_i, \qquad i = 1, \dots, 17.$$

---

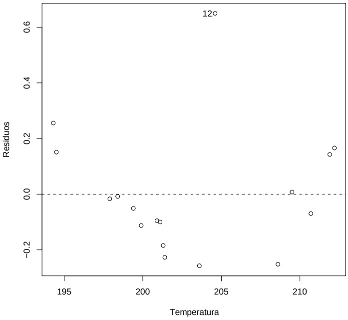
[6] El ambiente para cálculo estadístico **R** puede ser descargado desde CRAN: https://cran.r-project.org/

[7] Datos disponibles en la biblioteca alr4 para R.

Datos de Forbes



(a) recta ajustada



(b) residuos vs. ajuste

Datos de Forbes

Ahora consideramos el modelo

$$100 \times \log_{10}(\text{Presión}_i) = \beta_0 + \beta_1 \,\text{Temperatura}_i + \epsilon_i,$$

para $i = 1, \dots, 17$.

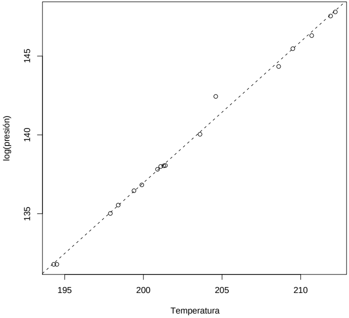
Se obtuvo:

$$\widehat{\boldsymbol{\beta}} = (-42.1378, 0.8955)^\top \qquad \text{y} \qquad s^2 = 0.1438$$

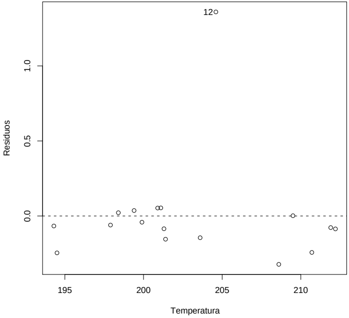
Además, $R^2 = 0.9950$.

Datos de Forbes

Recta de regresión y gráfico de residuos para los datos de Forbes[8].



(a) recta ajustada



(b) residuos vs. ajuste

[8] datos transformados

Datos de Huber (1981)[9]

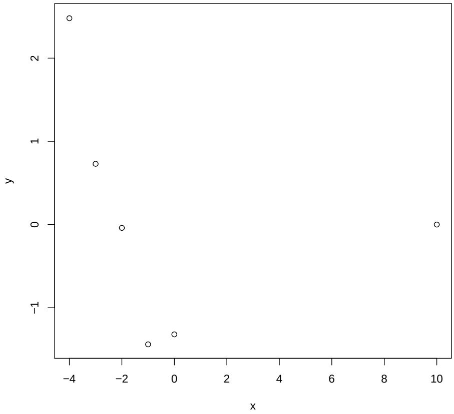
Considere el conjunto de datos hipotéticos de Huber.

Y	2.48	0.73	−0.04	−1.44	−1.32	0.00
x	−4	−3	−2	−1	0	10

[9] *Robust Statistics*. Wiley, New York

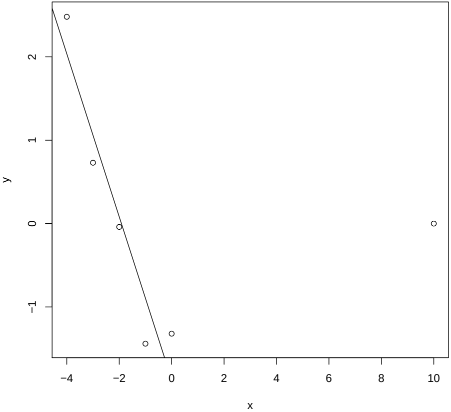
Datos de Huber

Diagrama de dispersión para los datos de Huber.



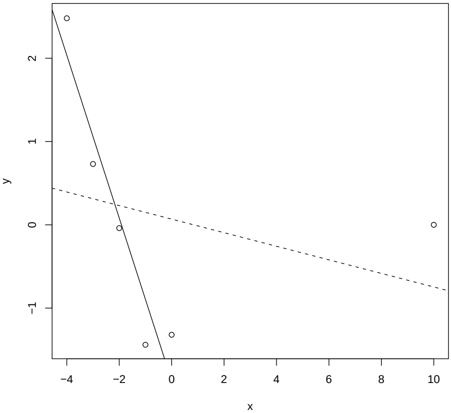
Datos de Huber

¿Qué opina de la recta de regresión?



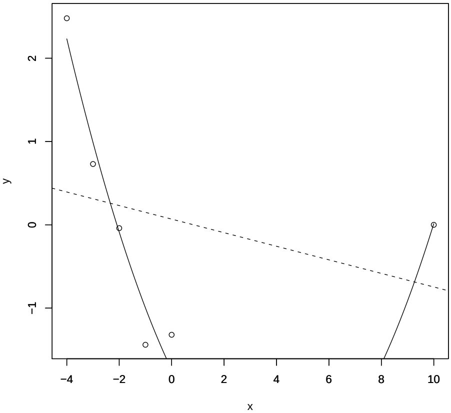
Datos de Huber

¿Y ahora?



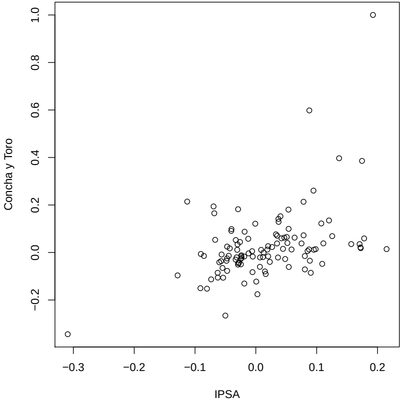
Datos de Huber

¿Cuál modelo prefiere?



Datos de Concha y Toro (Osorio y Galea, 2006)[10]

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



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

Modelo CAPM (Valoración de Activos de Capital), Sharpe (1964)[11]

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

usando datos observados, podemos escribir

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

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).

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

Datos de Concha y Toro

Modelo CAPM (Valoración de Activos de Capital), Sharpe (1964)[11]

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

usando datos observados, podemos escribir

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

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).

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

Modelo CAPM (Valoración de Activos de Capital), Sharpe (1964)[11]

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

usando datos observados, podemos escribir

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

Características del problema:

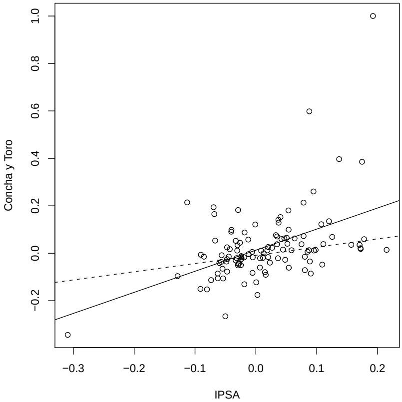
- Relación lineal entre las variables.
- Posibles periodos de alta volatilidad.

Hipótesis de interés:

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

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

Datos de Concha y Toro



Ajuste usando errores normales (—) y Cauchy (– –).

Cemento Portland (Woods, Steinour y Starke, 1932)[12]

Estudio experimental relacionando la emisión de calor durante la producción y endurecimiento de 13 muestras de cementos Portland. Woods et al. (1932) consideraron cuatro compuestos para los clinkers desde los que se produce el cemento.

La respuesta (Y) es la emisión de calor después de 180 días de curado, medido en calorías por gramo de cemento. Los regresores son los porcentajes de los cuatro compuestos: aluminato tricálcico (X_1), silicato tricálcico (X_2), ferrito aluminato tetracálcico (X_3) y silicato dicálcico (X_4).

[12] Industrial and Engineering Chemistry **24**, 1207-1214.

Cemento Portland (Woods, Steinour y Starke, 1932)

Y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Observación:

En efecto, existe una relación lineal aproximada, pues $x_1 + x_2 + x_3 + x_4 \approx 100$.

Datos de Puromycin (Treolar, 1974)

Modelo Michaelis-Menten: usado para el estudio de cinética de enzimas.

Permite estudiar la relación entre velocidad inicial de una reacción enzimática a la concentración de un substrato x a través de la ecuación:

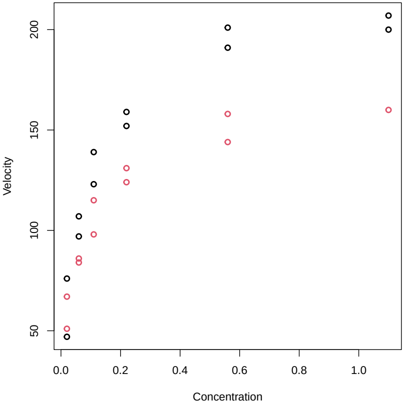
$$f(x, \boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x}, \qquad \boldsymbol{\beta} = (\beta_1, \beta_2)^\top.$$

Diferenciando f con relación a β_1 y β_2, obtenemos

$$\frac{\partial f}{\partial \beta_1} = \frac{x}{\beta_2 + x}, \qquad \frac{\partial f}{\partial \beta_2} = -\frac{\beta_1 x}{(\beta_2 + x)^2}.$$

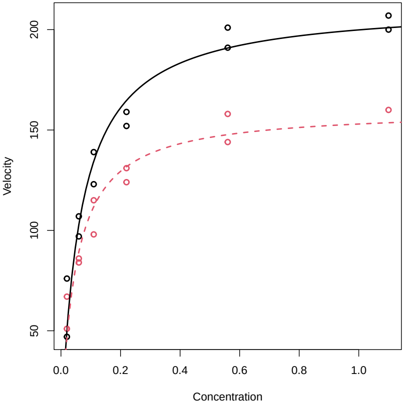
Datos de Puromycin

Velocidad de una reacción enzimática como función de la concentración del substrato para un experimento sobre enzimas tratadas con Puromycin.



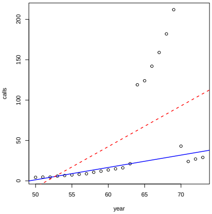
Datos de Puromycin

Velocidad de una reacción enzimática como función de la concentración del substrato para un experimento sobre enzimas tratadas con Puromycin.

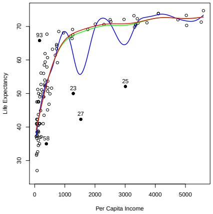


Modelos lineales son los *bloques de construcción* para metodologías más complejas, tales como:

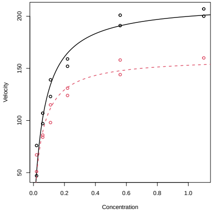
- Modelos lineales generalizados.
- Modelos no lineales.
- Modelos de regresión espacial.
- Regresión multivariada.
- Datos longitudinales, GMANOVA.
- Regresión semiparamétrica.
- Modelos con efectos mixtos.



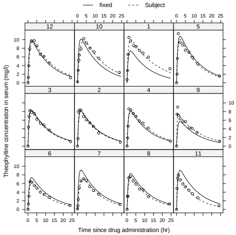
(a) regresión LAD



(b) splines penalizados



(c) regresión no-lineal



(d) modelos mixtos