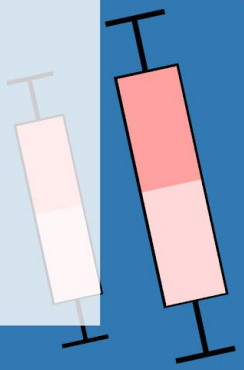
Herramientas Estadísticas y Forecast

Clase 8: Modelos de regresión lineal

Luis Gutiérrez – Ricardo Olea

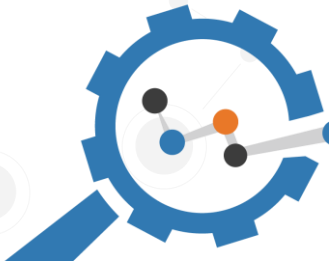
DATA SCIENCE UC
FACULTAD DE MATEMÁTICAS

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE



Contenido del curso

1. Introducción a la estadística y análisis descriptivo
2. Análisis descriptivo y gráfico
3. Probabilidad y Distribuciones
4. Muestreo
5. Inferencia estadística: Pruebas de hipótesis
6. Taller Práctico de inferencia
7. Introducción a los modelos estadísticos
8. Modelos predictivos I: Modelos de regresión lineal.
9. Modelos predictivos II: Regresión logística y otros modelos.
10. Modelos de Forecasting I.
11. Modelos de Forecasting II
12. Taller de Forecast



Modelo de Regresión Lineal

Interesa **pronosticar** mediante una combinación lineal de variables, preferentemente independientes entre ellas, una variable de interés Y.

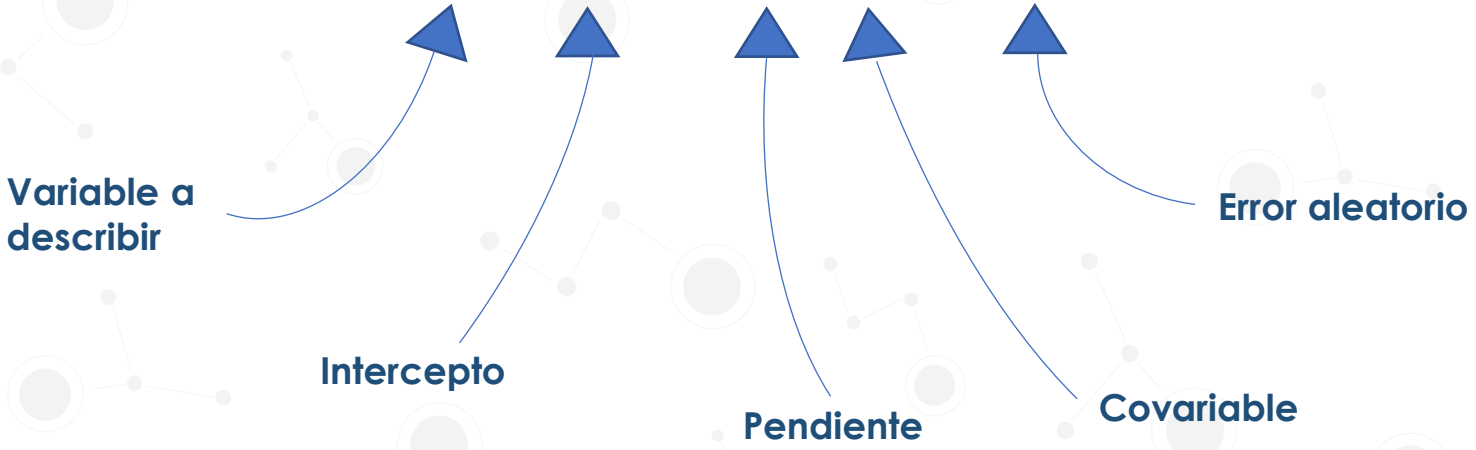
La recta o plano que forma está combinación lineal representan el valor esperado de Y, condicionado a la información $X_1, \dots, X_k$.

La estimación de los coeficientes que ponderan cada variable se obtienen mediante mínimos cuadrados.

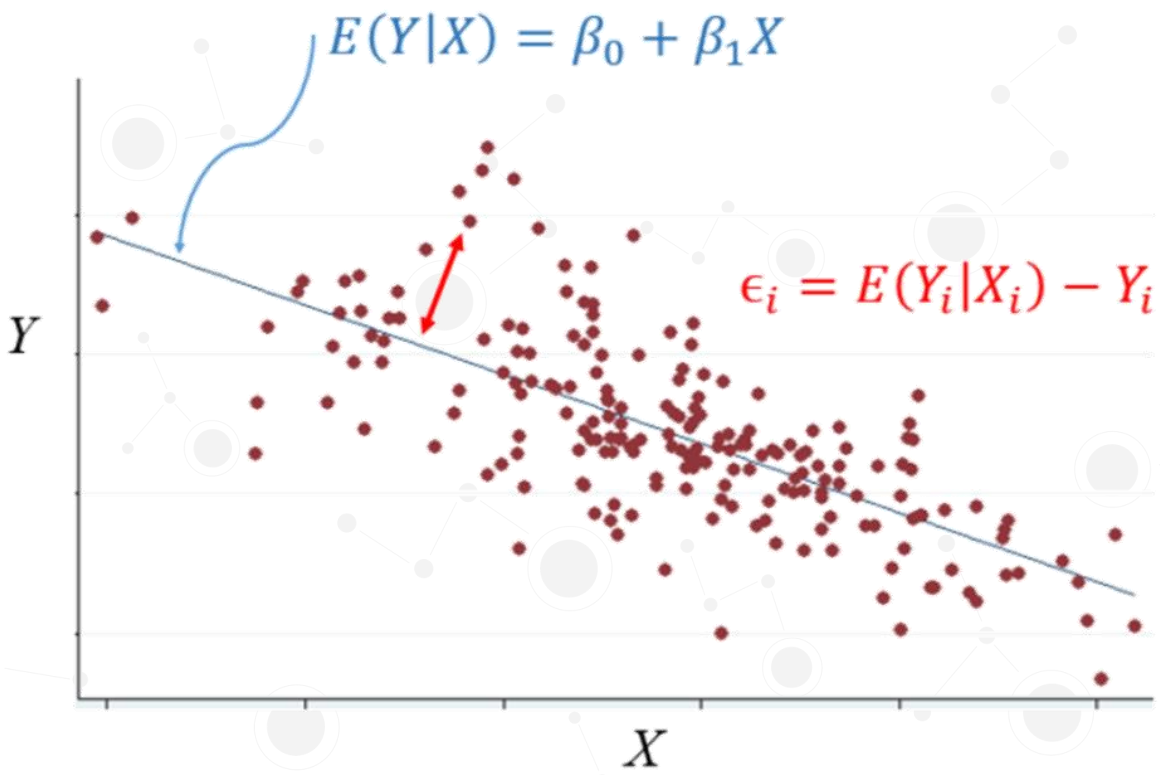
Modelo de Regresión Lineal

El caso más simple es denominado Modelo de **Regresión Lineal Simple**, que es cuando se utiliza una variable predictora (o covariable) para pronosticar una variable respuesta Y.

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



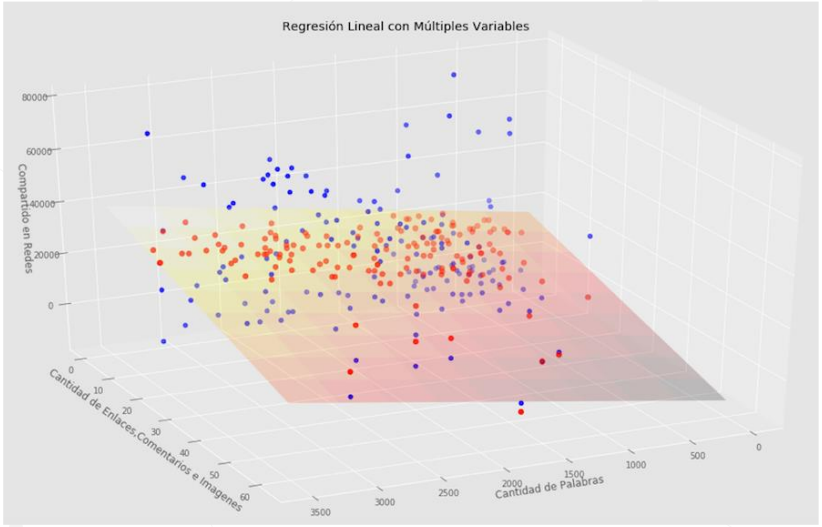
Modelo de Regresión Lineal



Modelo de Regresión Lineal

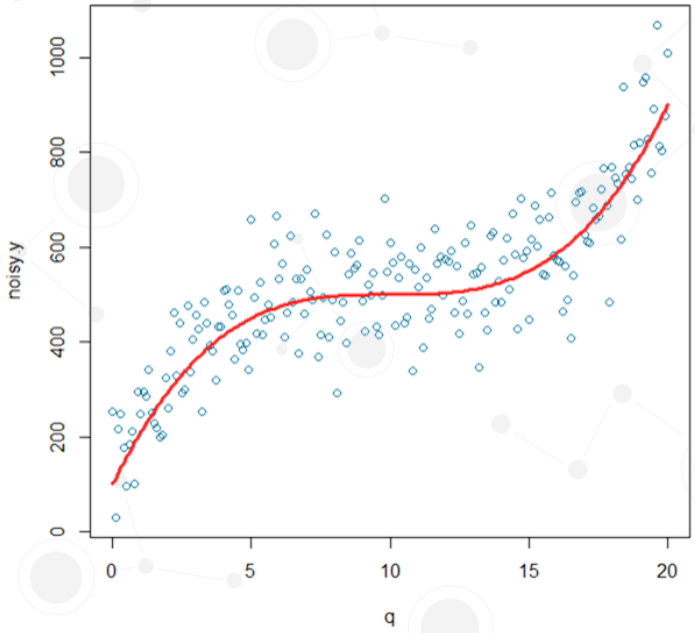
Un Modelo de **Regresión Lineal Múltiple**, corresponde cuando se utilizan varias covariables que intentan explicar una variable a pronosticar.

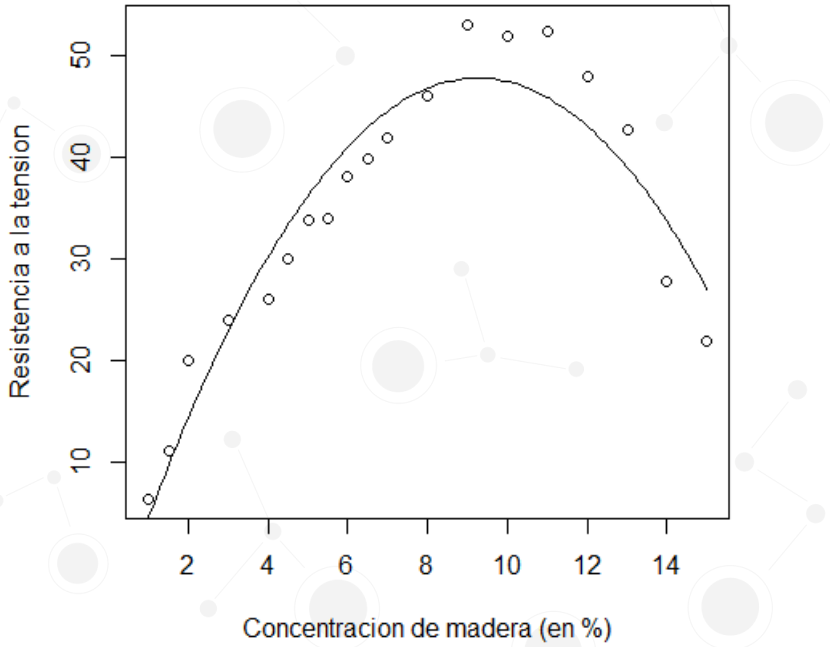
$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i$$



Modelo de Regresión Lineal

¿Por qué **Lineal**?





Modelo de Regresión Lineal

Dos preguntas importantes:

a) Probar si los parámetros obtenidos son significativos implica comprobar si la variable explica a la respuesta.

$$\mathrm{H}_0: \beta_i = 0 \quad \text{vs.} \quad \mathrm{H}_1: \beta_i \neq 0$$

Se realizará un test t individual para cada parámetro. Además, se puede obtener los intervalos de confianza de cada uno de ellos.

Modelo de Regresión Lineal

Dos preguntas importantes:

b) Probar si el modelo obtenidos es significativo implica comprobar si el modelo es correcto.

$$\mathrm{H}_0: \beta_0 = \beta_1 = \cdots = \beta_k = 0 \quad \text{vs.} \quad \mathrm{H}_1: \text{algún } \boldsymbol{i}, \boldsymbol{j} \text{ tal que } \beta_i \neq \beta_j$$

Se realizará un test F a partir de la tabla ANOVA del modelo. En este caso, se reemplaza la idea de Tratamiento por Regresión.

Modelo de Regresión Lineal

El **Coeficiente de Determinación** R^2 es una medida de qué también se ajusta la ecuación de regresión lineal múltiple a los datos muestrales. Un ajuste perfecto daría como resultado $R^2 = 1$, y un ajuste muy bueno da como resultado un valor cercano a 1. Un menor ajuste se relaciona con un valor de R^2 cercano a 0.

En **R** se puede realizar un modelo de regresión con la función **lm(Y~X1+X2+…Xk**). Para ver las pruebas de hipótesis y su R^2, se puede realizar un **summary()** del objeto anterior

Modelo de Regresión Lineal

Se desea predecir la esperanza de vida con algunas covariables

Residuals:

Min	1Q	Median	3Q	Max
-1.47095	-0.53464	-0.03701	0.57621	1.50683

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.103e+01	9.529e-01	74.542	< 2e-16 ***
habitantes	5.014e-05	2.512e-05	1.996	0.05201 .
asesinatos	-3.001e-01	3.661e-02	-8.199	1.77e-10 ***
ingresos	2.701e-04	3.087e-04	0.875	0.3867
universitarios	4.658e-02	1.483e-02	3.142	0.00297 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared: 0.736, Adjusted R-squared: 0.7126
F-statistic: 31.37 on 4 and 45 DF, p-value: 1.696e-12

Actividad

[Descargar los datos aquí]

Table 1.2 Munich rent index: description of variables including summary statistics

Variable	Description	Mean/ frequency in %	Std.- dev.	Min/max
rent	Net rent per month (in Euro)	459.43	195.66	40.51/1,843.38
rentsqm	Net rent per month per square meter (in Euro)	7.11	2.44	0.41/17.72
area	Living area in square meters	67.37	23.72	20/160
yearc	Year of construction	1,956.31	22.31	1918/1997

Actividad

location	Quality of location according to an expert assessment	
	1 = average location	58.21
	2 = good location	39.26
	3 = top location	2.53
bath	Quality of bathroom	
	0 = standard	93.80
	1 = premium	6.20
kitchen	Quality of kitchen	
	0 = standard	95.75
	1 = premium	4.25
cheating	Central heating	
	0 = without central heating	10.42
	1 = with central heating	89.58
district	District in Munich	

Actividad 1

Se pretende explicar y predecir el valor del arriendo por metro cuadrado en función del tamaño del departamento. Para ello, ajuste un modelo de regresión lineal simple considerando **rentsqm** como variable respuesta y **area** como predictor. Responda las siguientes preguntas:

1. ¿Tiene sentido realizar una regresión entre estas variables?
2. ¿Qué ocurre con el valor de los arriendos a medida que aumenta el tamaño del departamento?
3. ¿Interprete el valor del **Coeficiente de Determinación** R^2 ?

Modelo de Regresión Lineal

Para que el modelo cumpla las expectativas requeridas, debe cumplir tres criterios en sus errores:

(a) **Normalidad**: Se espera que los errores tengan distribución normal, con media 0 y cierta varianza.

(b) **Independencia**: Los errores deben ser independientes entre si.

(c) **Homocedasticidad**: Los errores deben tener una dispersión constante en cualquier instante de la variable.

Modelo de Regresión Lineal

Se puede aplicar una prueba de hipótesis en cada caso, donde se busca rechazar la hipótesis nula (valor-p alto)

(a) Normalidad: Test de Shapiro-Wilk
En R: **shapiro.test()**

(b) Independencia: Test de Durbin-Watson
En R: **dwtest()**

(c) Homocedasticidad: Test de Breusche-Pagan
En R: **bptest()**

Modelo de Regresión Lineal

Un indicador para elegir un modelo corresponde al **Criterio de Información de Akaike (AIC)**. El AIC es una medida de calidad basada en la bondad de ajuste. Si se poseen varios modelos, se optaría por aquel con valor mínimo de AIC.

Un método computacional para elegir un modelo es a través de comparar AIC. La función **stepAIC()** del paquete **MASS** permite seleccionar el método mediante el argumento **direction = c("both", "backward", "forward"),** que significa en ambas direcciones, ir retirando variables, o ir agregando variables, respectivamente.

Actividad 2

a) Proponga un modelo de regresión lineal múltiple para el valor del arriendo por metro cuadrado. No considere en el modelo la variable **distric**.

b) Para el modelo encontrado, valide los supuestos.

Modelo de Regresión Lineal

Los modelos estadísticos tienen como fin determinar si una o más variables pueden entregar información para explicar una variable respuesta.

Por lo tanto, si se conoce una nueva observación para el conjunto de covariables, se puede realizar un **pronóstico** de la variable respuesta.

La función **predict()** realiza una estimación puntual, junto con un intervalo de confianza para la nueva observación.

Actividad 3

1. Realice una predicción del valor promedio del arriendo por metro cuadrado para un departamento de 30 metros cuadrados con las siguientes características: Año 1990, top location, quality of bathroom premium, quality of kitchen premium, with central heating.
2. Realice una predicción del valor promedio del arriendo por metro cuadrado para un departamento con las mismas características que en 1. pero con una superficie 110 metros cuadrados.
3. Basado en 1. y 2. en qué tipo de departamento recomienda invertir, grandes o chicos?

Taller

Es de interés realizar un modelo de regresión lineal para poder modelar la esperanza de vida basado en información de inmunización, factores mortalidad, factores económicos y factores sociales, además, de otros factores relacionados a la salud. Para ello dispone de la esperanza de vida de distintos países entre los años 2000 y 2005, la base fue conformada con información de la OMS y las Naciones Unidas.

(**Descargue los datos aquí**)

Nos interesa responder las siguientes preguntas:

1. ¿Cuáles son las variables predictoras que realmente afectan a la esperanza de vida?
2. ¿Cómo afecta la tasa de mortalidad de lactantes y adultos a la esperanza de vida?
3. ¿Cuál es el impacto de la escolaridad en la esperanza de vida?
4. ¿Cuál es el impacto de la cobertura de inmunización en la esperanza de vida?

Ajuste un modelo de regresión lineal múltiple, con un método automatizado de dirección backward, utilizando solo variables significativas al 10, no considere el country al general el modelo.