Universidad del Rosario

REGRESIÓN

Dora Suárez
Juan Fernando Pérez

Otras Pruebas de Hipótesis

Parámetro – Caso	Hipótesis Nula	Comando R	¿Cuándo se usa?
Coeficiente de correlación de Pearson	$\rho = 0$	`cor.test(x, y, method = “pearson")`	- Las dos variables son continuas
Coeficiente de correlación de Spearman	$\rho = 0$	`cor.test(x, y, method = "spearman")`	- Las dos variables son continuas
Independencia de variables categóricas	$\rho = 0$	`chisq.test(table(x,y))`	- Las dos variables son cuantitativas

Teorema del límite central

Sea $\{X_i\}$, i=1...,N, una secuencia de variables aleatorias iid con $E(X) = \mu$ y $V(X) = \sigma^2$. Entonces:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \rightarrow N(0,1)$$

Método de Monte-carlo

Simular una variable aleatoria exponencial
Calcular su promedio
Repetir muchas veces el experimento
Obtener estadísticas de la distribución del promedio

Introducción

Modelos

Una teoría o hipótesis a menudo predice una relación entre dos variables. ¿Cómo evalúan los científicos si los datos apoyan o refutan una relación ?

Video

Población
Fenómeno

Muestreo

Muestra
Datos

- Corte transversal — Tiempo Fijo
- Series temporales — Unidad Fija
- Panel / Longitudinal — Ni tiempo ni unidad fija

Modelo Estadístico
Función de representación de los parámetros de la población

Inferencia
Tipos / Formas

- Estimación puntual
- Estimación por intervalos
- Test de hipótesis

Estadística descriptiva
Descripción o caracterización de los datos a través de cantidades numéricas

Definiciones

Modelo: Representación simplificada de la realidad que contiene los aspectos mas importantes de la misma

Modelo estadístico: Modelo que incorpora un elemento de aleatoriedad

Parámetros: Cantidades fijas y usualmente desconocidas que indexa el modelo y representan características de la población

Análisis de regresión

Un modelo de regresión es un modelo estadístico en que $\boxed{\text{alguna característica}}$ distribucional de la $\boxed{\text{variable de interés}}$ es afectada por $\boxed{\text{otras variables.}}$

Media (Gauss)
Mediana (Laplace)

Y : Variable de interés
Variable respuesta
Variable dependiente
Regresando

Hecho importante
X afecta el comportamiento de Y

X : Variable independiente
Regresor
Variable exploratoria
Covariables

Regresión lineal simple

Caso más importante

Si la media de Y es afectada por X tenemos entonces que:

$$\mathrm{E(Y)} = \mu = f(x)$$

El caso más importante ocurre cuando la función de x es lineal

$$\boxed{\mu = \beta_0 + \beta_1 x}$$

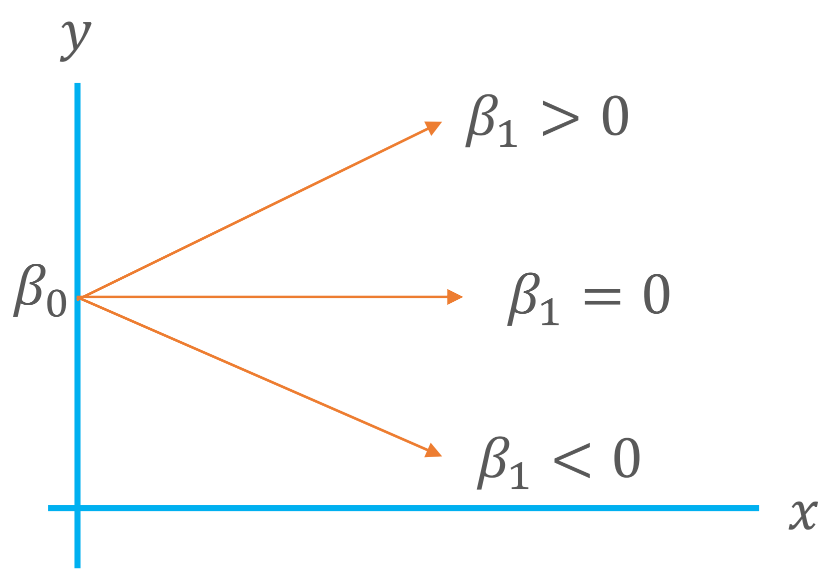
Modelo indexado por parámetros desconocidos y fijos

Modelo de regresión lineal simple

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

$Y \sim alguna\ distribución(\beta_0 + \beta_1 x, \sigma^2)$

$\beta_1, \beta_2, \sigma^2$: Parámetros



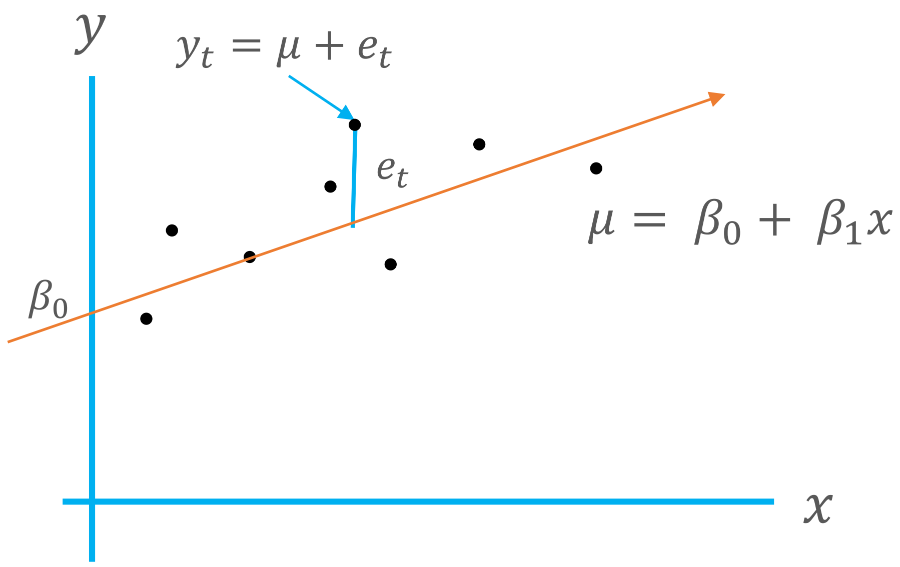
Objetivo:
A partir de los datos hacer inferencia sobre los parámetros

Modelo de regresión lineal simple

$$\mu_t = \beta_0 + \beta_1 x_t \longleftrightarrow y_t = \beta_0 + \beta_1 x_t + e_t$$

t = 1, 2,..., N
N: Tamaño de la muestra

Error: No observable



Ejemplo – Ley de gas ideal

Ley: A temperatura constante, el volumen de una masa fija de gas es inversamente proporcional a la presión que este ejerce. Matemáticamente se puede expresar así:

$$PV = nRT$$

Relación Lineal

Si la presión es constante, entre más temperatura, mayor presión

n: Moles de gas
R: Constante universal de los gases ideales
V: Volumen
P: Presión
T: Temperatura

Ejemplo – Ley de gas ideal

Ley: A temperatura constante, el volumen de una masa fija de gas es inversamente proporcional a la presión que este ejerce. Matemáticamente se puede expresar así:

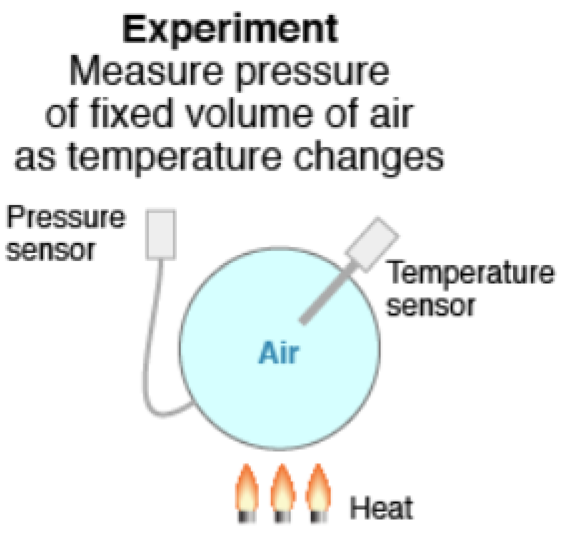
$$PV = nRT$$

Relación Lineal

Si la presión es constante, entre más temperatura, mayor presión

n: Moles de gas
R: Constante universal de los gases ideales
V: Volumen
P: Presión
T: Temperatura

Experimento



Data

Temperature (ºC)	Pressure (Pa)
20	111
22	111
25	106
33	112
44	117
47	122
59	123
70	128

```
experimento <- data.frame(temperatura = c(20, 22, 25, 33, 44, 47, 59, 70),  presion = c(111,111,106, 112, 117, 122, 123, 128))

p <- ggplot(experimento, aes(x = temperatura, y = presion))
      p + geom_point(shape = 20, colour = "blue") +
       labs(title = "Ley Gases",
          x = "Temperatura",
          y = "Presión",
          subtitle = "Resultados del experimento")
```

¿Qué es considerado lineal?

El modelo de regresión debe ser lineal en los parámetros, es decir en las cantidades desconocidas. Para el modelo lineal simple los parámetros desconocidos son β_0 y β_1

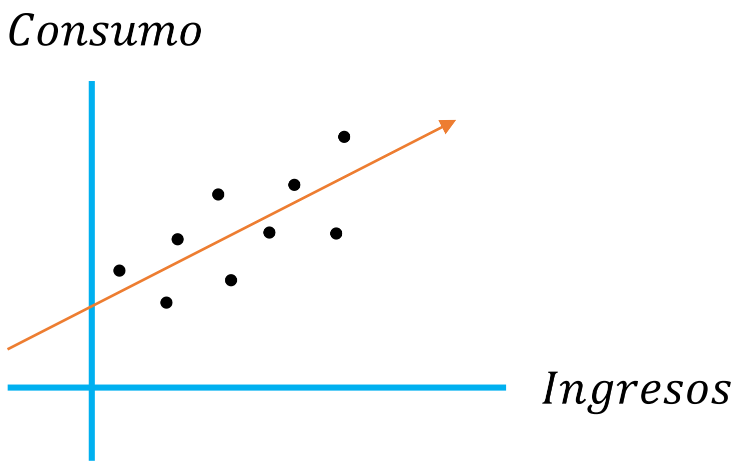
$$y_t = \beta_0 + \beta_1 x_t + e_t \qquad y_t = \beta_0 + \beta_1 x_t^2 + e_t \qquad y_t = \beta_0 + \beta_1 \ln(x_t) + e_t$$

$$\ln(y_t) = \beta_0 + \beta_1 x_t + e_t \qquad y_t = \beta_0 + \beta_1^2 x_t + e_t \qquad y_t = \beta_0 + \ln(\beta_1) x_t + e_t$$

¿Qué representan los parámetros?

β_0 **Intercepto**
Media de y cuando x = o
* Preguntarse si tiene sentido su interpretación dependiendo del problema

β_1 **Inclinación**
Variación en la media de la variable respuesta cuando la regresora aumenta una unidad



$$presion_t = \beta_0 + \beta_1 temperatura_t + e_t$$

β_0 Presión del gas dentro de la bomba cuando la temperatura es O°C

β_1 Aumento promedio de la presión cuando la temperatura aumenta un grado centígrado

Estimación por mínimos cuadrados

[Objetivo] Encontrar estimadores de los parámetros del modelo de regresión lineal simple.

[¿Cómo?] Minimizando cosas "malas" o maximizando cosas "buenas"

Malo: Errores positivos y negativos se cancelan $\sum_{t=1}^{N} e_t$

Regresión L1 (Difícil de estimar por esta vía) $\sum_{t=1}^{N} |e_t|$

Regresión L2 (Estimación de mínimos cuadrados) $\sum_{t=1}^{N} e_t^2$

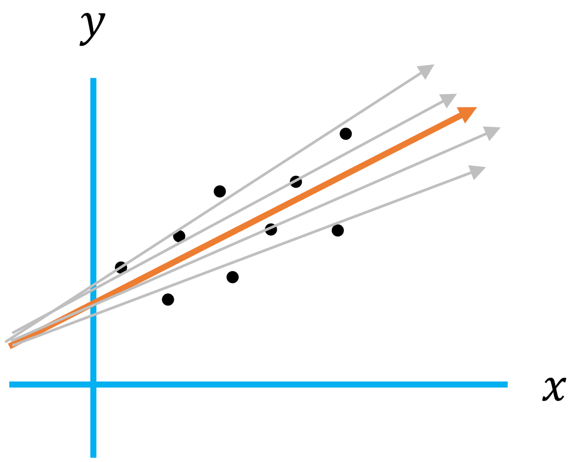
Errores
Minimizar los errores o una medida agregada de los errores

Verosimilitud
Máximizar la probabilidad de que la muestra en conjunto sea la observada

Estimación por mínimos cuadrados

En la estimación por mínimos cuadrados se buscan los argumentos que minimicen la suma de los cuadrados de los errores

$$\sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} (y_t - (\beta_0 + \beta_1 x))^2$$



Obtención de los estimadores

1) Desarrollar el cuadrado
2) Derivar respecto a β_0 y β_1 e igualar a cero
3) Despejar los valores de β_0 y β_1 del sistema de ecuaciones obtenido en 2.
4) Usar $\hat{\beta}_0$ y $\hat{\beta}_1$ como estimadores para calcular la media a predecir

$\mu = \hat{\beta}_0 + \hat{\beta}_1 x$

El procedimiento completo puede ser encontrado en [Ref 1 – pag 540]

Estimadores de mínimos cuadrados

Intercepto	Pendiente
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\hat{\beta}_1 = \frac{N\sum x_t y_t - \sum x_t \sum y_t}{N\sum x_t^2 - (\sum x_t)^2}$
$var(\hat{\beta}_0) = \sigma^2 \frac{\sum x_t^2}{N\sum (x_t - \bar{x})^2}$	$var(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum (x_t - \bar{x})^2}$
$cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{\sum (x_t - \bar{x})^2}$	

Cuanto mayor sea σ^2, menos precisos son los estimadores

A mayor tamaño de muestra, las varianzas de los estimadores tienden a cero (Consistentes)

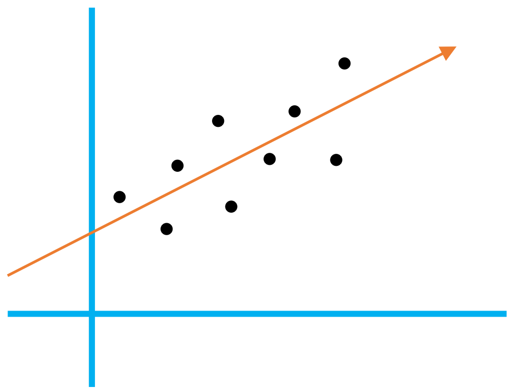
Entre más variables los valores de x, más precisos son los estimadores

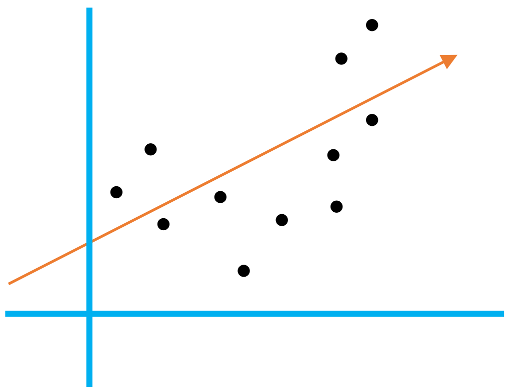
Estimación por mínimos cuadrados

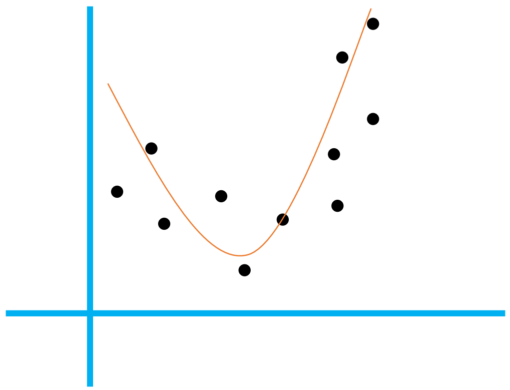
```
x = experimento$temperatura
y = experimento$presion
n = length(x)

beta1 = (n*sum(x*y)-sum(x)*sum(y))/(n*sum(x^2)-sum(x)^2)
beta0 = mean(y)-beta1*mean(x)

Beta1 = mean(experimento$presion)
modelo <- lm(presion ~ temperatura, data = experimento)
summary(modelo)
plot(modelo)
```

Supuestos del modelo de regresión lineal

[1] El modelo propuesto ofrece una buena descripción de la realidad







```
library(lmtest)
resettest(modelo, power = 2, type = "regressor")
```

Implicación
El modelo es adecuado para describir el fenómeno de interés

Descomposición del error

$$Y'Y = \hat{Y}\hat{Y} + \hat{e}'\hat{e}$$

Suma de cuadrados
de los errores

(SST)

$$\boxed{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \boxed{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} + \boxed{\sum_{i=1}^{n}\hat{e}_i^2}$$

Suma de los errores **no**
explicados por la regresión

SSE

Suma de cuadrados de los errores explicados
por la regresión

(SSR)

$$SST = SSR + SSE$$

Prueba F: Si la varianza de la regresión y del total son casi lo mismo, entonces el modelo es significativo

Coeficiente de determinación

$$\frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$$1 = \boxed{\frac{SSR}{SST}} + \frac{SSE}{SST}$$

Coeficiente de determinación R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

% de variabilidad explicado por la regresión

Propiedades:

- ✓ Es **no** decreciente respecto al numero de variables regresoras
- ✓ Es una proporción (toma valores entre 0 y 1)
- ✓ Mide la variación proporcional en la variación total de y que es explicada por el modelo

Coeficiente de determinación ajustado

$$\overline{\mathrm{R}}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Su finalidad es medir el poder de discriminación de un modelo

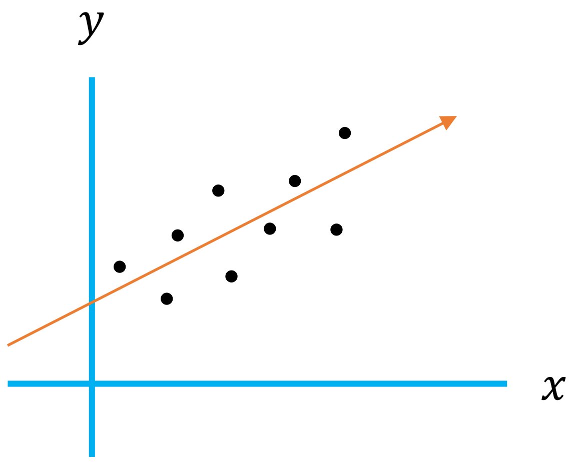
Propiedades:

- ✓ Puede decrecer cuando aumenta la cantidad de variables regresoras
- ✓ No es una proporción (puede tomar valores negativos)
- ✓ No se interpreta, sólo se usa como criterio de decisión entre diferentes modelos

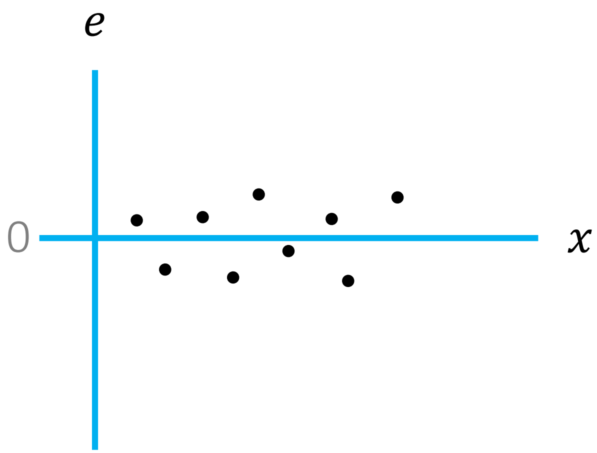
Supuestos del modelo de regresión lineal

[2] La media de los errores es cero

t.test(modelo$residuals)



$$y_t - \mu = e_t$$



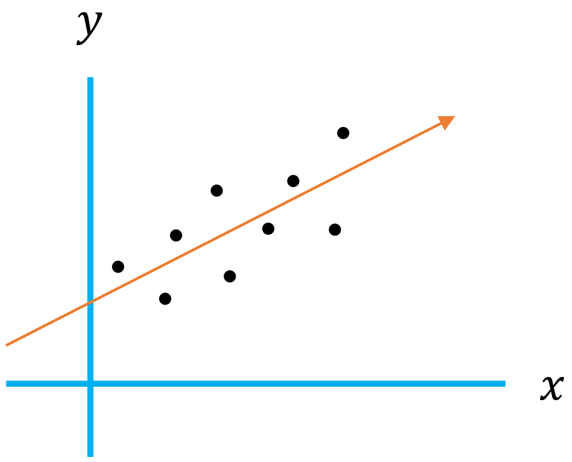
Implicación:
El valor esperado de la variable respuesta es efectivamente la media
*Para x fija

$$\mathrm{E(y_t)} = \mu_t = \beta_0 + \beta_1 x_t$$

Supuestos del modelo de regresión lineal

[3] La varianza de los errores es constante

bptest(modelo, studentize = T)

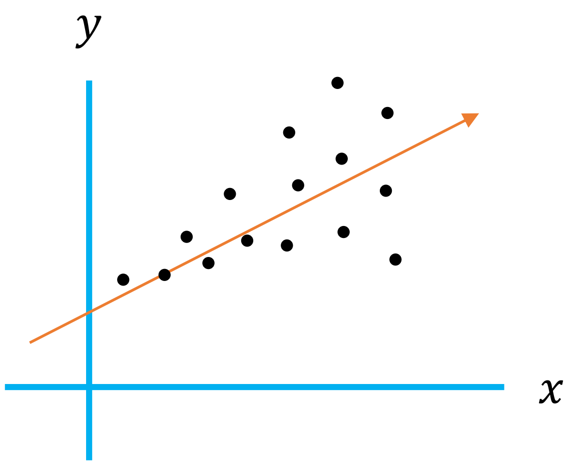


Homocedástico

$var(e_t) = \sigma^2$

Implicación

La varianza de la variable respuesta es también fija

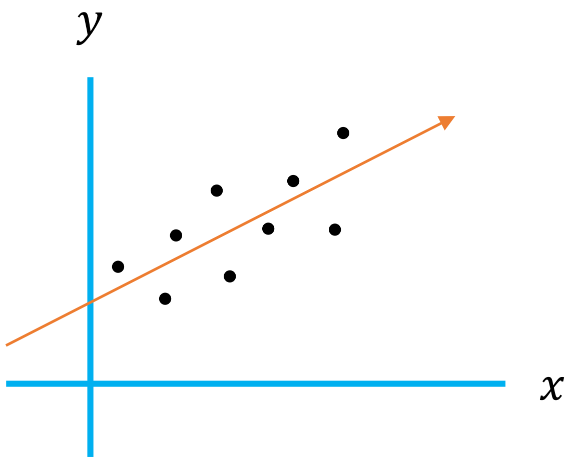


Heterocedástico

$var(e_t) = \sigma_t^2$

Supuestos del modelo de regresión lineal

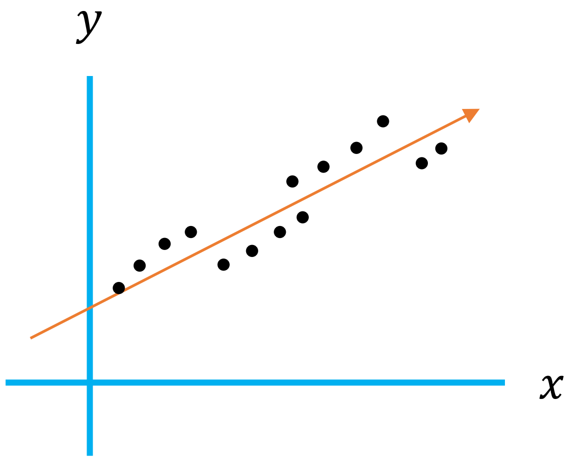
[4] No hay covarianza entre errores diferentes



$$cov(e_i, e_j) = 0;\ i \neq j$$

Implicación

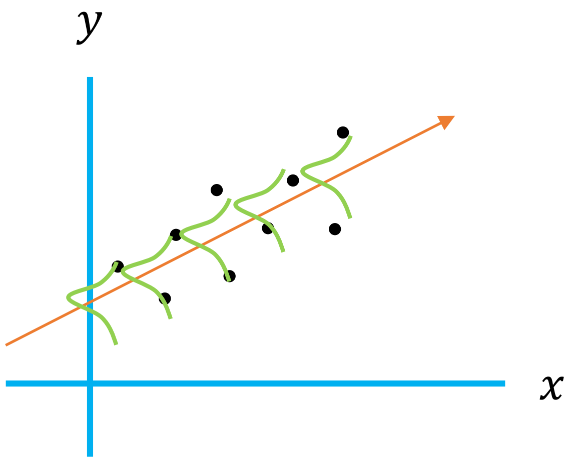
El comportamiento de una de las observaciones no influye en el comportamiento de las otras



$$cov(e_i, e_j) \neq 0;\ i \neq j$$

Supuestos del modelo de regresión lineal

[5] (Dependiendo del método de estimación) Los errores tienen distribución normal



Implicación

Permite construir intervalos de confianza en caso de que la estimación sea intervalar

```
shapiro.test(modelo, studentize = T)
```

Breve recordatorio de Álgebra Lineal

Matrices y sus “signos”

Sea A una matriz simétrica de dimensión $n \times n$ y sea z un vector de $n \times 1$. Acerca de A decimos que es:

Matriz
Definida Positiva

$$z'Az > 0$$

Matriz
Definida Negativa

$$z'Az < 0$$

Matriz
Semi Definida Positiva

$$z'Az \geq 0 \quad \forall z \neq 0$$

Matriz
Semi Definida Negativa

$$z'Az \leq 0 \forall z \neq 0$$

Si $A - B$ es definida positiva, entonces

$$z'Az > z'Bz \qquad \forall z$$

$$tr(A) > tr(B)$$

Derivada de Matrices

Sea A una matriz de dimensión $n \times n$, z un vector de $n \times 1$ y sea z un vector de $n \times 1$. Tenemos que:

[1] $$\frac{\partial z'w}{\partial z} = w$$

[2] $$\frac{\partial z'Az}{\partial z} = (A + A')z$$

Si A es simétrica:

$$\frac{\partial z'Az}{\partial z} = 2Az$$

Rangos de Matrices

El rango de una matriz es el número de filas y columnas linealmente independientes. Si A es una matriz simétrica invertible de dimensión $m \times m$ y sea B de dimensión $m \times p$, entonces:

[1] $B'AB$

es simétrica para cualquier matriz B

[2] Si $rango(B) = m$, entonces $\text{rango}(B'AB) = m$

[3] Si A es definida positiva y $rango(B) = m$, entonces $B'AB$ es definida positiva

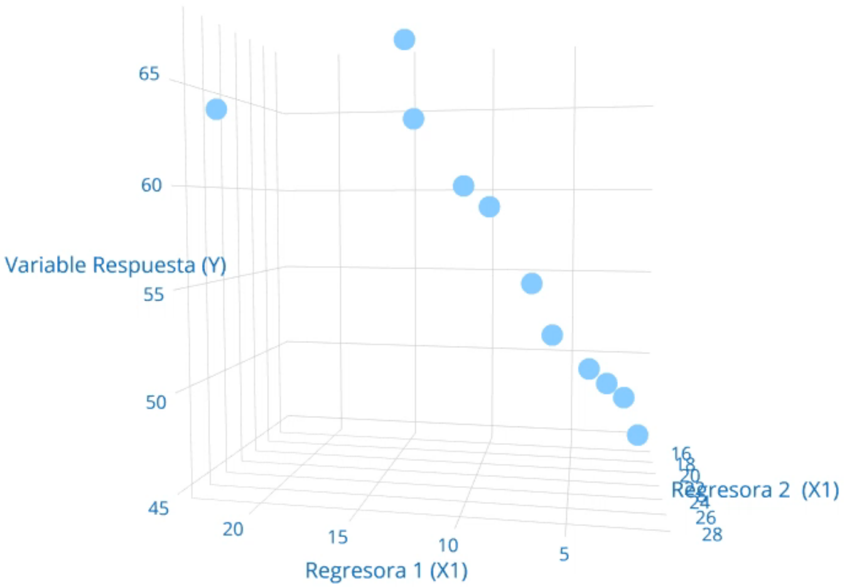
Regresión Lineal Múltiple

Regresión lineal múltiple

¿Qué pasa si tenemos más de una variable explicativa?

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e_t$$

Ejemplo de Regresión Múltiple (3 variables)



Interpretación de los parámetros:

β_0: Media de Y cuando todos los regresores son cero

β_j: Es la variación en la media de Y cuando x_j aumenta una unidad y todas las demás variables regresores quedan constantes

Modelo

Observación	y	x_1	x_2	…	x_p
1	y_1	x_{11}	x_{12}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…	…
N	y_N	x_{N1}	x_{N2}	…	x_{Np}

La inversión en medios publicitarios es uno de los costos más recurrentes de las empresas. Se desea estimar el valor de las ventas con base en las cantidades invertidas en los diferentes medios publicitarios.

Modelo

Observación	y	x_1	x_2	…	x_p
1	y_1	x_{11}	x_{12}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…	…
N	y_N	x_{N1}	x_{N2}	…	x_{Np}

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times p+1} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{bmatrix}_{p+1 \times 1} \quad e = \begin{bmatrix} e_0 \\ e_1 \\ \cdots \\ e_n \end{bmatrix}_{n \times 1}$$

$$Y = \boldsymbol{X}\beta + e$$

Objetivo:
Estimar β

Supuestos

[s1] El modelo propuesto ofrece una buena descripción de la realidad

[s2] $\mathrm{E}(e) = 0$

[s3] $\mathrm{V}(e) = \sigma^2$

[s4] $\mathrm{cov}(e_t, e_s) = 0$

[s5] Las columnas de **X** son linealmente independientes (Rango completo)

[s6] Los errores tienen distribución normal

Estimación – Mínimos cuadrados

Estimación de mínimos cuadrados ordinarios MCO

$$Y = \boldsymbol{X}\beta + e \longrightarrow \hat{\beta} = \underset{\beta \in R^p}{\text{arg\_min}}\ e'e \longrightarrow$$

Cuadrados de los errores

$$e'e = (Y - \boldsymbol{X}\beta)'(Y - \boldsymbol{X}\beta)$$
$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta = 0 \longrightarrow \hat{\beta} = \boxed{(X'X)^{-1}}(X'Y)$$

Por [s5] esta matriz existe

Propiedades del estimador MCO

[4] Es insesgado,

$$E(\hat{\beta}) = \beta$$

[2] Es un estimador lineal

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

(Puede ser escrito de la forma Ay')

[3] Matriz de covarianzas

$$cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

$$\widehat{cov}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$$

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-p} \longrightarrow \hat{e} = Y - \boldsymbol{X}'\beta$$

Propiedades del estimador MCO

[4] Bajo normalidad,

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$$

[5] $\hat{\sigma}^2$ es un estimador consistente

[6] Bajo los primeros 4 supuestos $\hat{\beta}$ es el mejor estimador lineal insesgado de β

[7] $\hat{\sigma}^2$ y $\hat{\beta}$ son independientes

Descomposición del error

$$Y'Y = \hat{Y}\hat{Y} + \hat{e}'\hat{e}$$

Suma de cuadrados
de los errores

(SST)

$$\boxed{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \boxed{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} + \boxed{\sum_{i=1}^{n}\hat{e}_i^2}$$

Suma de los errores **no**
explicados por la regresión

SSE

Suma de cuadrados de los errores explicados
por la regresión

(SSR)

$$SST = SSR + SSE$$

Coeficiente de determinación

$$\frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$$1 = \boxed{\frac{SSR}{SST}} + \frac{SSE}{SST}$$

Coeficiente de determinación R^2

$$\mathrm{R}^2 = 1 - \frac{SSE}{SST}$$

% de variabilidad explicado por la regresión

Propiedades:

- ✓ Es **no** decreciente respecto al numero de variables regresoras
- ✓ Es una proporción (toma valores entre 0 y 1)
- ✓ Mide la variación proporcional en la variación total de y que es explicada por el modelo

Coeficiente de determinación ajustado

$$\overline{R}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Su finalidad es medir el poder de discriminación de un modelo

Propiedades:

- ✓ Puede decrecer cuando aumenta la cantidad de variables regresoras
- ✓ No es una proporción (puede tomar valores negativos)
- ✓ No se interpreta, sólo se usa como criterio de decisión entre diferentes modelos

Práctica en R

```
rm(list=ls)
modelo <- lm(iris$Sepal.Length ~ iris$Petal.Width + iris$Petal.Length)
summary(modelo)
plot(modelo)
```