

REGRESIÓN

Dora Suárez
Juan Fernando Pérez

Otras Pruebas de Hipótesis

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

Teorema del límite central

Sea $\{X_i\}$, $i=1,\dots,N$, 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

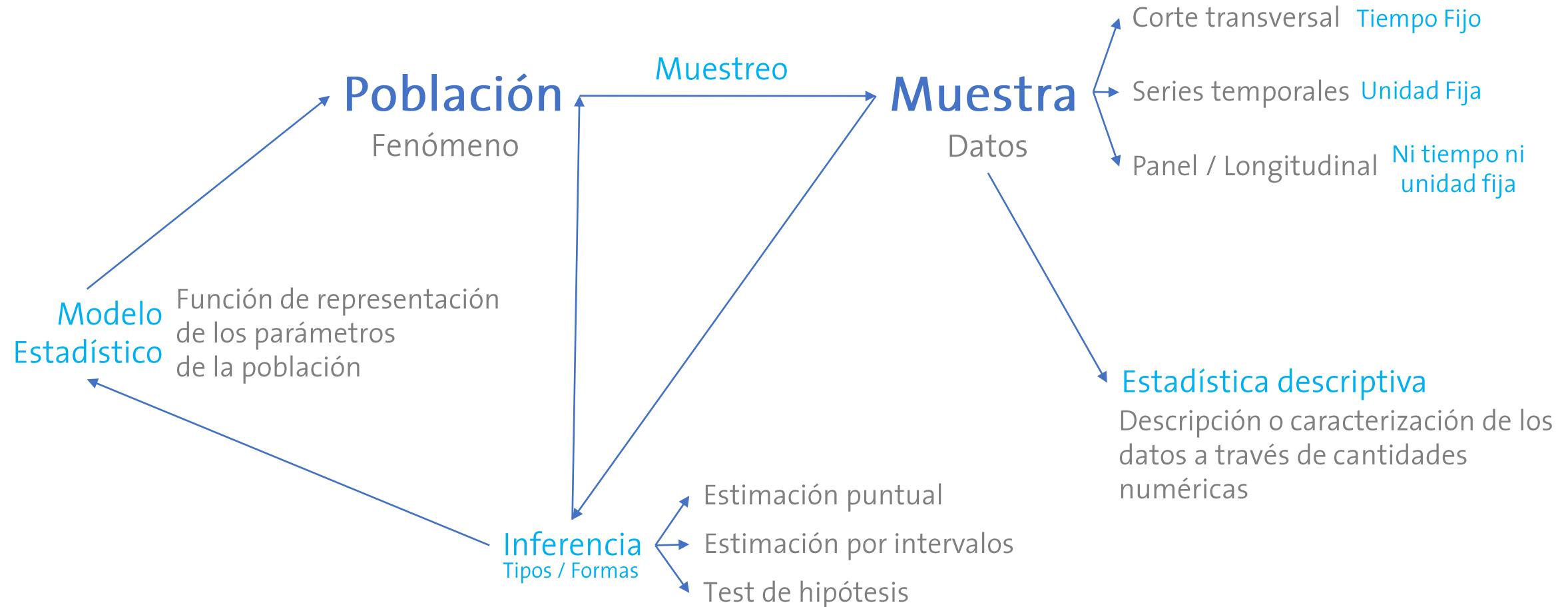
Obtener estadísticas de la distribución del promedio

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](#)



Definiciones

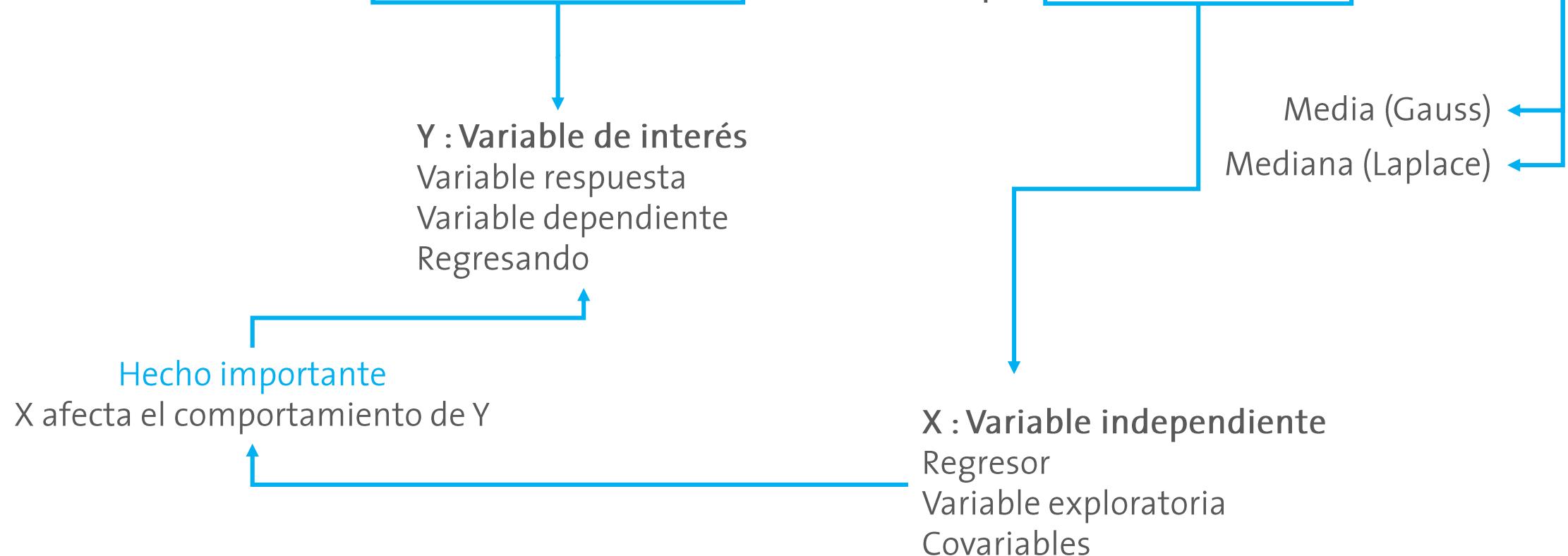
Modelo: Representación simplificada de la realidad que contiene los aspectos mas importantes de la misma

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 **alguna característica distribucional** de la **variable de interés** es afectada por **otras variables**.



Regresión lineal simple

Caso más importante

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

$$E(Y) = \mu = f(x)$$

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

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

Modelo indexado por
parámetros desconocidos y fijos

Modelo de regresión lineal simple

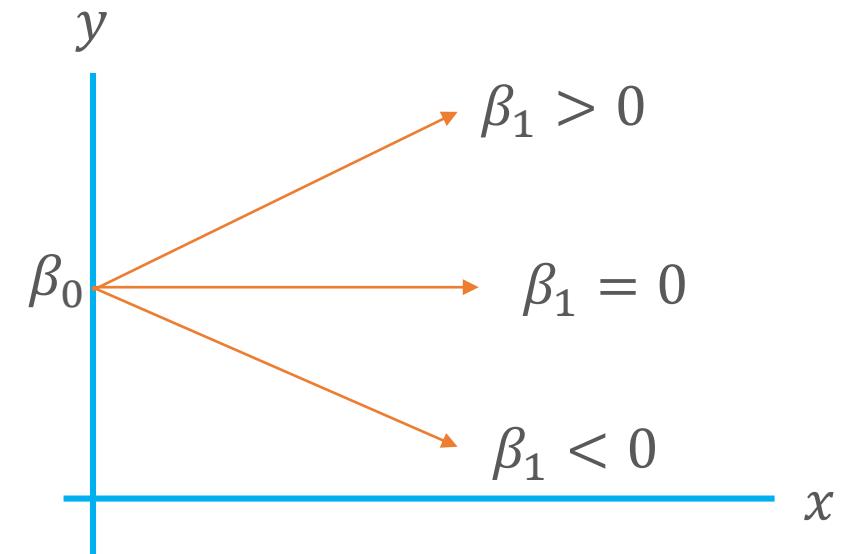
$$\mu = \beta_0 + \beta_1 x$$

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

$\beta_0, \beta_1, \sigma^2$: Parámetros

Objetivo:

A partir de los datos hacer
inferencia sobre los parámetros



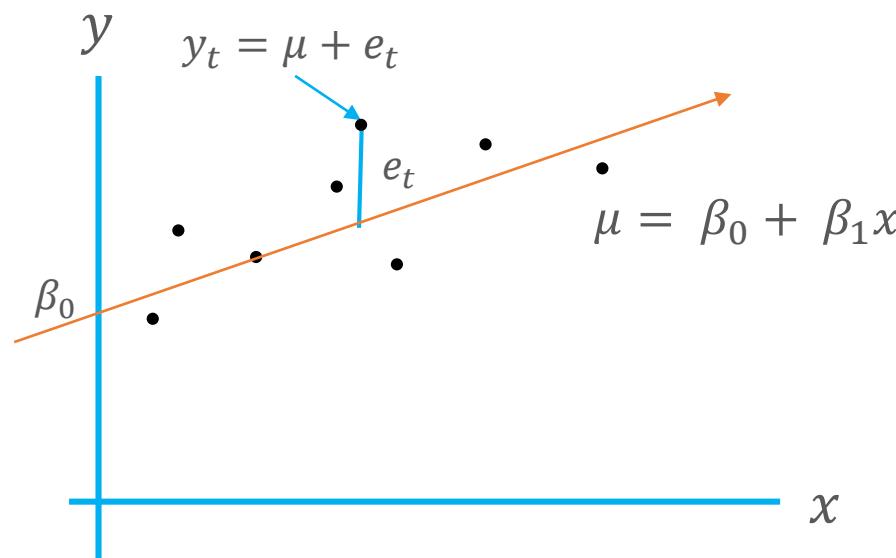
Modelo de regresión lineal simple

$$\mu_t = \beta_0 + \beta_1 x_t \quad \longleftrightarrow \quad y_t = \beta_0 + \beta_1 x_t + e_t \quad t = 1, 2, \dots, N$$

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

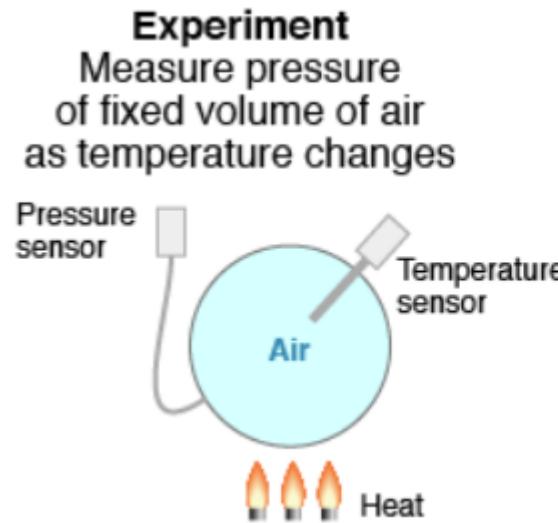
R: Constante universal de los gases ideales

V: Volumen

P: Presión

T: Temperatura

Experimento



Data

Temperature (°C)	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

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

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

¿Qué representan los parámetros?

 β_0

Intercepto

Media de y cuando x = 0

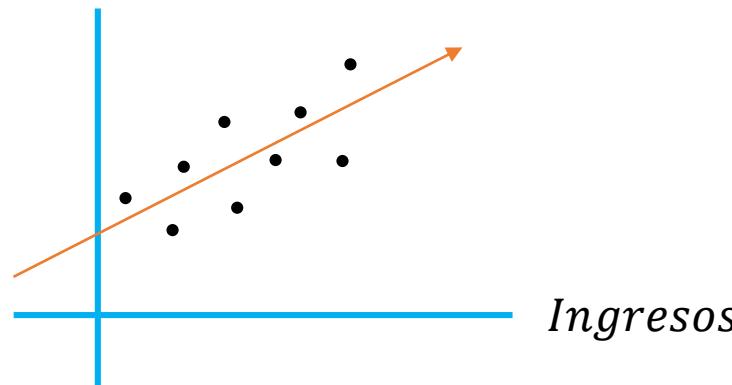
* Preguntarse si tiene sentido su interpretación dependiendo del problema

 β_1

Inclinación

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

Consumo



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

 β_0

Presión del gas dentro de la bomba cuando la temperatura es 0°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

Regresión L1 (Difícil de estimar por esta vía)

Regresión L2 (Estimación de mínimos cuadrados)

$$\sum_{t=1}^N e_t$$

$$\sum_{t=1}^N |e_t|$$

$$\sum_{t=1}^N e_t^2$$

Minimizar los errores o una medida agregada de los errores

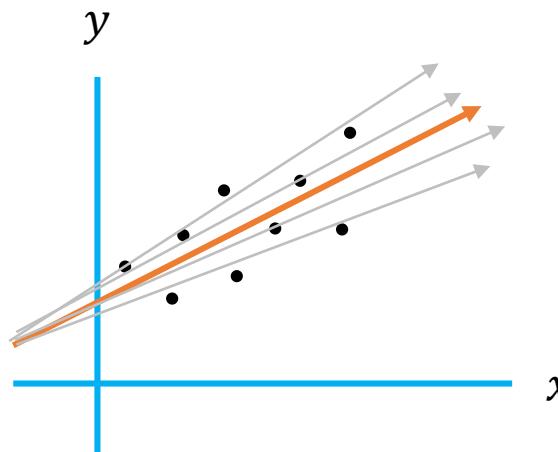
Errores

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)
```

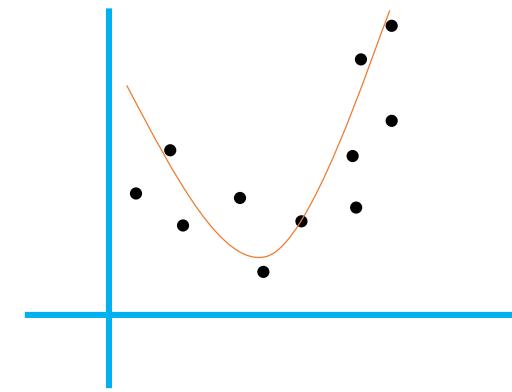
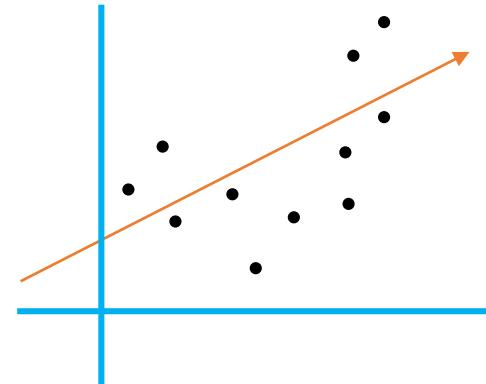
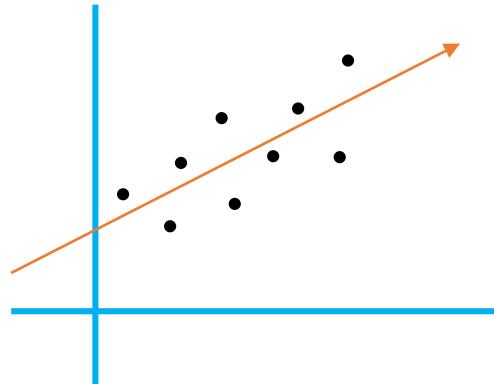
```
modelo <- lm(presion ~ temperatura, data = experimento)
```

```
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)

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

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

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

Su finalidad es medir el poder de discriminación de un modelo

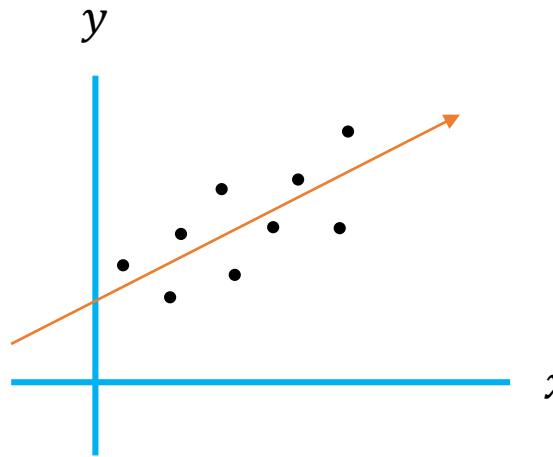
Propiedades:

- ✓ 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

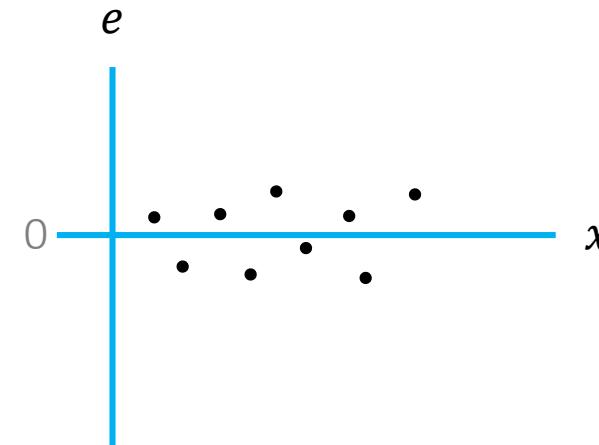
Supuestos del modelo de regresión lineal

[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

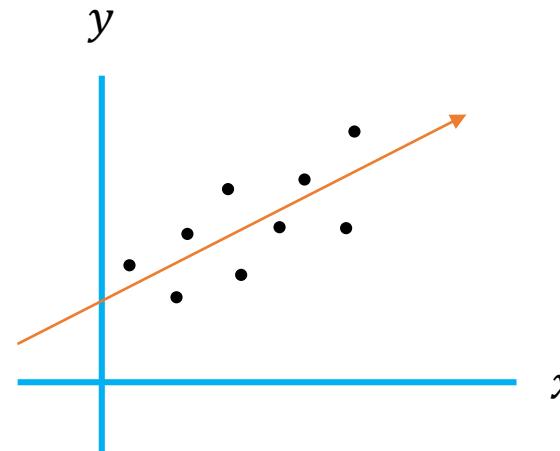
*Para x fija

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

Supuestos del modelo de regresión lineal

[3] La varianza de los errores es constante

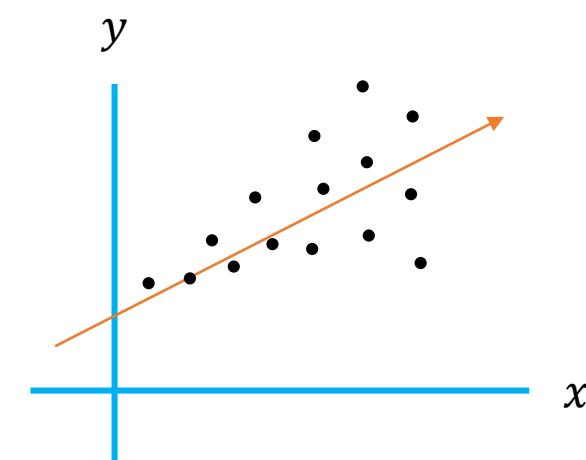
bptest(modelo, studentize = T)



Homocedástico

$$\text{var}(e_t) = \sigma^2$$

Implicación
La varianza de la variable respuesta es también fija

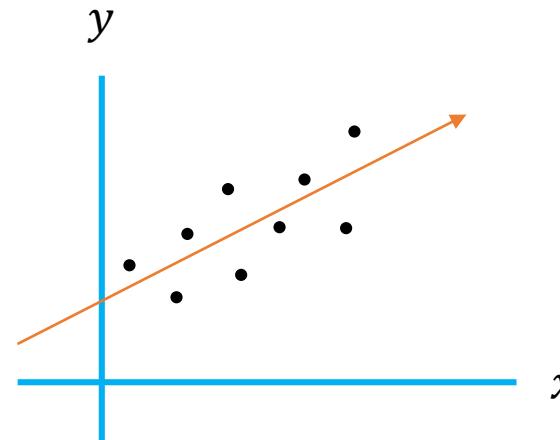


Heterocedástico

$$\text{var}(e_t) = \sigma_t^2$$

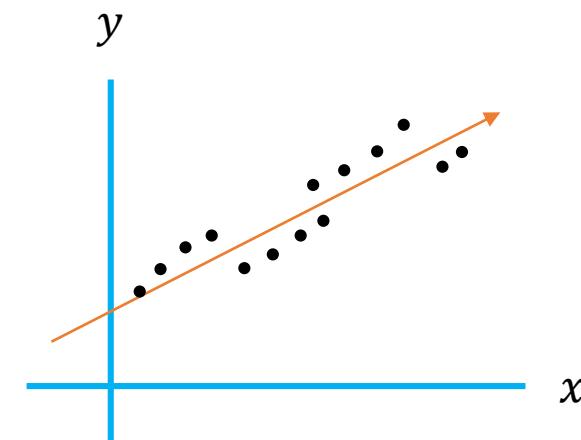
Supuestos del modelo de regresión lineal

[4] No hay covarianza entre errores diferentes



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

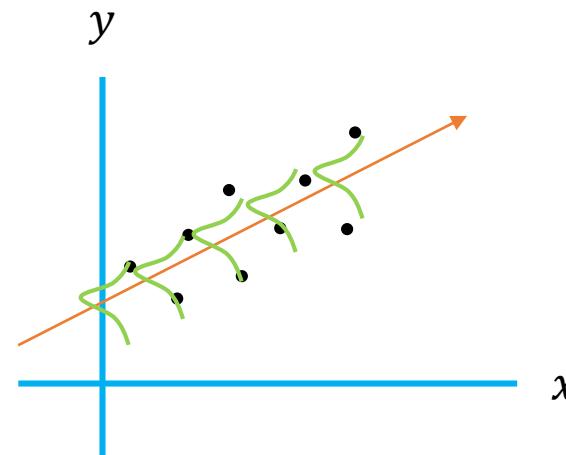
Implicación
El comportamiento de una de las observaciones no influye en el comportamiento de las otras



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

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

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

Si $A - B$ es definida positiva, entonces

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

$$\text{tr}(A) > \text{tr}(B)$$

Derivada de Matrices

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

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

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

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 $\text{rango}(B) = m$, entonces
 $\text{rango}(B'AB) = m$

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

Regresión Lineal Múltiple

Regresión lineal múltiple

¿Qué pasa si tenemos más de una variable explicativa?

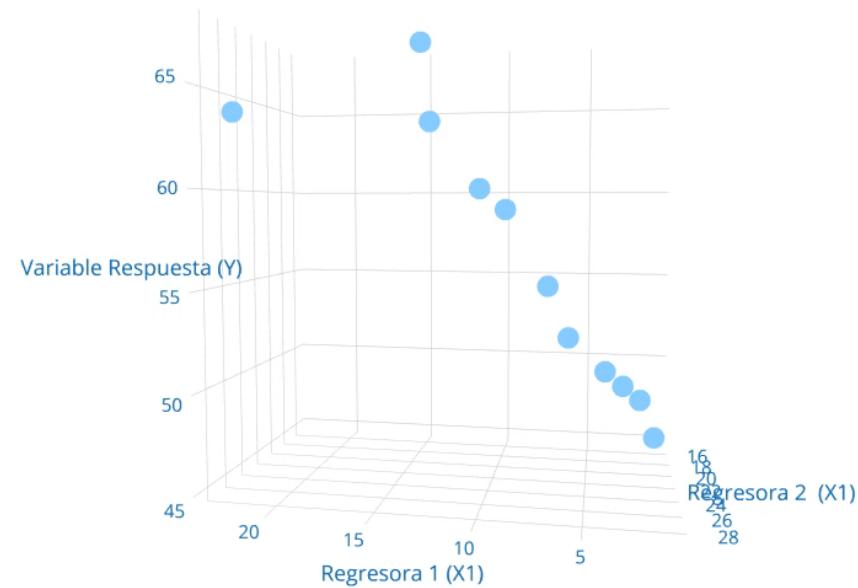
$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e_t$$

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

Ejemplo de Regresión Múltiple (3 variables)



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}

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

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

$$Y = X\beta + e$$

Objetivo:
Estimar β

Supuestos

- [s1] El modelo propuesto ofrece una buena descripción de la realidad
- [s2] $E(e) = 0$
- [s3] $V(e) = \sigma^2$
- [s4] $\text{cov}(e_t, e_s) = 0$
- [s5] Las columnas de \mathbf{X} son linealmente independientes (Rango completo)
- [s6] Los errores tienen distribución normal

Estimación – Mínimos cuadrados

Estimación de mínimos cuadrados ordinarios MCO

$$Y = X\beta + e \longrightarrow \hat{\beta} = \arg \min_{\beta \in R^p} e'e \longrightarrow \text{Cuadrados de los errores}$$

$$\begin{aligned} e'e &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \end{aligned}$$

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta = 0 \longrightarrow \hat{\beta} = \boxed{(X'X)^{-1}}(X'Y) \xrightarrow{\text{Por [s5] esta matriz existe}}$$

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

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

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

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

Propiedades del estimador MCO

[4] Bajo normalidad,

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

[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)

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

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

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

↓
% 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

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

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)
```

Modelos lineales con relaciones logarítmicas

Otra forma de ajustar modelos es dada por:

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

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

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

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

Cuando las relaciones entre x y y no son lineales, es posible aplicar transformaciones que permitan llegar a una relación lineal.

El caso más común son las transformaciones logarítmicas, al realizar estas transformaciones, las interpretaciones cambian.

Modelos lineales con relaciones logarítmicas

Otra forma de ajustar modelos es dada por:

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

β_1 : Incremento de unidades en “y” cuando aumenta 1 unidad en “x”

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

$\beta_1/100$: Incremento en unidades de “y” cuando aumenta un 1% en “x”

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

$\beta_1 * 100$: Incremento porcentual de “y” cuando aumenta 1 unidad en “x”

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

β_1 : Incremento porcentual de “y” cuando aumenta un 1% en “x”

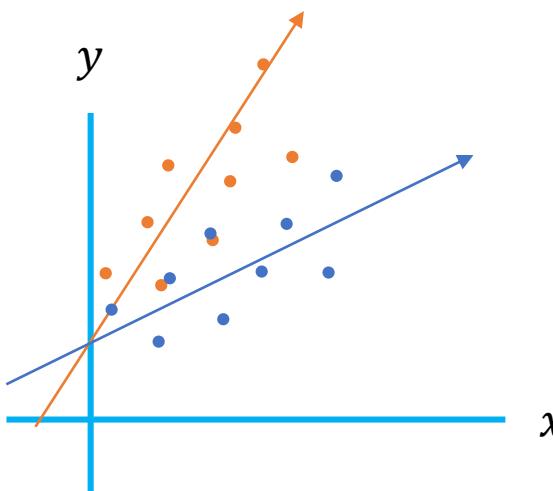
Regresión Lineal ante la presencia de variables categóricas

¿Qué pasa si una de las variables de estudio es categórica?

Para poder incluir en el modelo una variable categórica dentro una regresión y poder medir su efecto se hace uso de variables dummy

Una variable dummy, es una variable que puede tomar únicamente dos valores, típicamente cero o uno. Por ejemplo el género de una persona:

$$\delta = \begin{cases} 1 & \text{si es Mujer} \\ 0 & \text{si es hombre} \end{cases}$$



¿Cómo incluir las variables dummy en una regresión?

$$\hat{y}_i = \beta_0 + (\beta_1 + \gamma\delta)x_i$$

Por ejemplo, si queremos ver el efecto en el consumo medio de las personas de acuerdo a sus ingresos y género:

$$consumo_{esperado} = \beta_0 + (\beta_1 + \gamma\delta)ingresos \quad \delta = \begin{cases} 1 & \text{si es Mujer} \\ 0 & \text{si es hombre} \end{cases}$$

$$\begin{aligned} Si \delta = 0 \Rightarrow & Hombres \\ & \beta_0 + \beta_1 \text{ingresos} \end{aligned}$$

$$\begin{aligned} Si \delta = 1 \Rightarrow & Mujeres \\ & \beta_0 + \beta_1 + (\beta_1 + \gamma) \text{ingresos} \end{aligned}$$

¿Cómo interpretamos los coeficientes obtenidos?

β_0 : Intercepto

β_1 : Aumento esperado en el consumo de los hombres por cada 1000 pesos adicionales en sus ingresos (suponiendo que los ingresos están en miles de pesos)

γ : Aumento/disminución esperada en el consumo de las mujeres respecto a los hombres

$\beta_1 + \gamma$: Aumento esperado en el consumo de las mujeres por cada 1000 pesos adicionales en sus ingresos (suponiendo que los ingresos están en miles de pesos)

¿Qué hacer si la variable categórica tiene más de dos categorías?

Rta: Si se cuenta con k categorías, crear k-1 variables dummy.

En ese caso, se debe establecer una categoría base o de comparación, frente a la que se van a comparar las demás. Por ejemplo, si se tienen las categorías: Hombre, Mujer, Otro:

$$\delta_1 = \begin{cases} 1 & \text{si es Mujer} \\ 0 & \text{si es hombre} \end{cases}$$

$$\delta_2 = \begin{cases} 1 & \text{si es Otro} \\ 0 & \text{si es hombre} \end{cases}$$

Las interpretaciones son por tanto, respecto a la diferencia entre hombres y las demás categorías

Ejemplo

Las Naciones Unidas disponen la información de 213 países de:

Region: África, Asia, Caribe, Europa, Latinoamérica, Norte América, Atlántico norte, Oceanía

Grupo: África, miembros de la OECD, otros

Tasa de fertilidad: Número de hijos por mujer

Ppgdp: Ingreso per-cápita en dólares

Expectativa de vida: Número de años esperados de vida en las mujeres

Mortalidad infantil: Número de niños que fallecen antes de 1 año por cada 1000 nacidos vivos

¿Existe diferencias en la expectativa de vida de las personas de áfrica respecto a las demás?

```
datos <- UN
datos1 <- datos %>% na.omit() %>% within(., group <- relevel(group,
ref="africa"))
levels(datos1$group)

modelo1 <- lm(lifeExpF ~ group , data = datos1)
summary(modelo1)

modelo2 <- lm(lifeExpF ~ infantMortality + group , data = datos1)
summary(modelo2)
```

Modelos para respuestas binarias

Modelos con respuesta binaria

¿Qué pasa si la variable respuesta es una variable binaria?

En este caso, la variable “y” puede tomar los valores cero o uno.

Supongamos que:

$$y_i = \begin{cases} 1 & \text{con probabilidad } p_i \\ 0 & \text{con probabilidad } 1 - p_i \end{cases}$$

$$E(y_i) = p_i$$

Si x toma valores en un rango diferente a cero o uno este modelo no es adecuado

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

Regresion Logistica

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i = z_i$$

$$p_i = \frac{e^{z_i}}{(1 + e^{z_i})}$$

Ejemplo, se desea estimar la probabilidad de que una persona no pague su cuenta de banco en función del balance y los ingresos de la misma

Regresion Logistica

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i = z_i \quad p_i = \frac{e^{z_i}}{(1+e^{z_i})}$$

Razón de chances (odss):

$$\frac{p_i}{1-p_i}$$

Por ejemplo, si la probabilidad de que ocurra un evento es del 80% entonces se espera que haya 4 veces más probabilidad de que ocurra que de que NO ocurra

$$\frac{0.8}{0.2} = 4$$

Regresion Logistica

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i = z_i$$

$$p_i = \frac{e^{z_i}}{(1 + e^{z_i})}$$

Los odds y el logaritmo de odds cumplen que:

Si $p(\text{verdadero}) = p(\text{falso})$, entonces $\text{odds}(\text{verdadero}) = 1$

Si $p(\text{verdadero}) < p(\text{falso})$, entonces $\text{odds}(\text{verdadero}) < 1$

Si $p(\text{verdadero}) > p(\text{falso})$, entonces $\text{odds}(\text{verdadero}) > 1$

A diferencia de la probabilidad que no puede exceder el 1, los odds no tienen límite superior.

Si $\text{odds}(\text{verdadero}) = 1$, entonces $\text{logit}(p) = 0$

Si $\text{odds}(\text{verdadero}) < 1$, entonces $\text{logit}(p) < 0$

Si $\text{odds}(\text{verdadero}) > 1$, entonces $\text{logit}(p) > 0$

La transformación logit no existe para $p = 0$

Muestras de test y muestras de entrenamiento

Las muestras de test son las muestras con las que vamos a verificar si el modelo tiene o no un buen desempeño

Las muestras de entrenamiento son las muestras con las que vamos a ajustar el modelo

Se suele tomar un 70% de los datos para entrenar el modelo y un 30% para probar su desempeño

Ajuste del modelo

Ajustar el modelo con los datos de entrenamiento, con la opción de enlace logístico en R

Establecer el punto de corte de la probabilidad (a partir de que punto se dice que es uno o cero)

Una vez estos datos estén ajustados, predecir la probabilidad en los datos de prueba

Matriz de confusión

Tabla de contingencia definida como:

		Datos de Prueba		P'
Predicción	Positivos	Positivos	Negativos	
	Positivos	Verdaderos Positivos	Falsos Positivos	N'
	Negativos	Falsos Negativos	Verdaderos Negativos	N'
		P	N	TOTAL

Sensibilidad

Razón de verdaderos positivos:

$$VPR = \frac{VP}{P} = \frac{VP}{VP + FN}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Ratio

Razón de falsos positivos:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + VN}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Accuracy - Exactitud

$$ACC = \frac{VP + VN}{P + N}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Especificidad

$$SPS = \frac{VN}{N} = \frac{VN}{FP + VN} = 1 - FPR$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Valor predictivo positivo

$$PPV = \frac{VP}{VP + FP}$$

		Datos de Prueba		
Predicciones	Positivos	Positivos	Negativos	
		VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Valor predictivo negativo

$$NPV = \frac{VN}{VN + FN}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Curva ROC

Gráfica de la especificidad vs la sensibilidad a medida que varía el umbral de discriminación

A		
VP=63	FP=28	91
FN=37	VN=72	109

100 100 200

VPR = 0.63

FPR = 0.28

ACC = 0.68

B		
VP=77	FP=77	154
FN=23	VN=23	46

100 100 200

VPR = 0.77

FPR = 0.77

ACC = 0.50

C		
VP=24	FP=88	112
FN=76	VN=12	88

100 100 200

VPR = 0.24

FPR = 0.88

ACC = 0.18

C'		
VP=76	FP=12	88
FN=24	VN=88	112

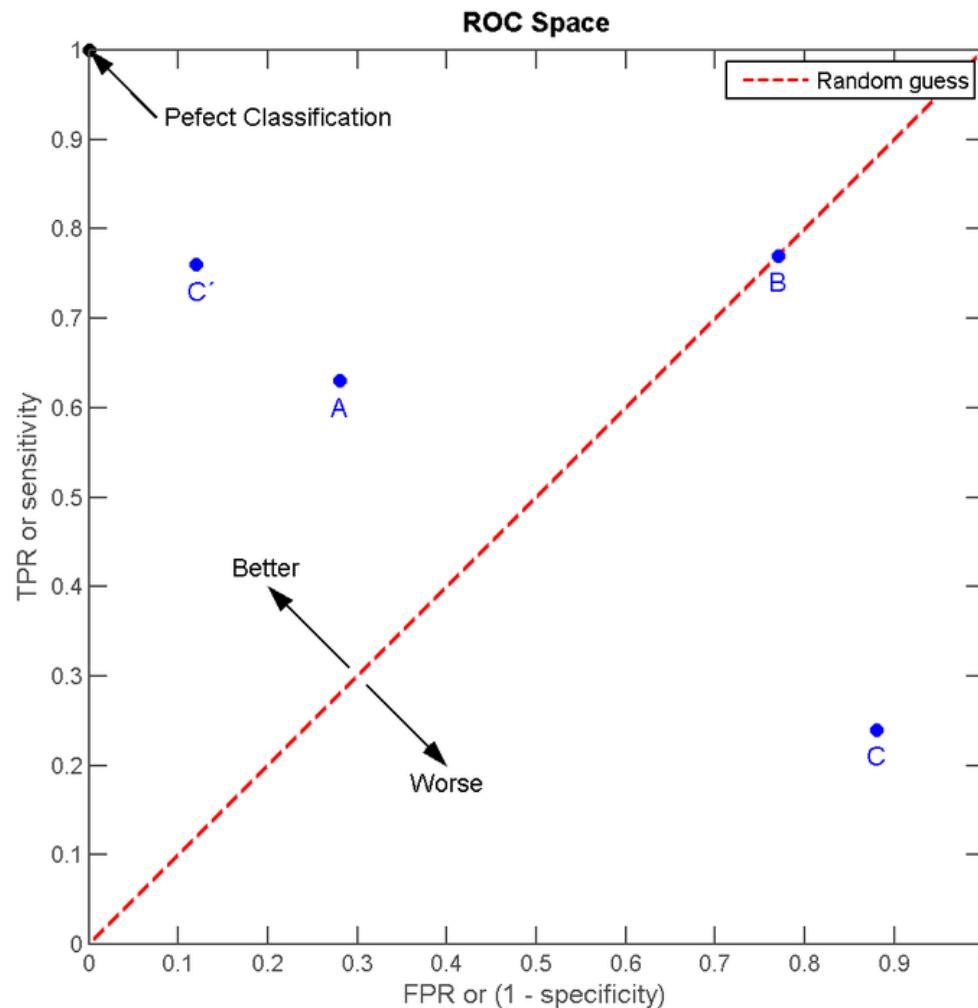
100 100 200

VPR = 0.76

FPR = 0.12

ACC = 0.82

Curva ROC



Curva ROC

Dibujar la curva ROC consiste en poner juntos todos los puntos correspondientes a todos los umbrales o puntos de corte, de tal modo que ese conjunto de puntos se parecerá más o menos a una curva en el espacio cuadrado entre (0,0) y (1,1)

Area bajo la curva: Este índice se puede interpretar como la probabilidad de que un clasificador ordenará una instancia positiva elegida aleatoriamente más alta que una negativa. Cuanto mayor sea este valor, mejor es nuestra regresión

Curva ROC

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

[0.5]: Es como lanzar una moneda.

[0.5, 0.6): Test malo.

[0.6, 0.75): Test regular.

[0.75, 0.9): Test bueno.

[0.9, 0.97): Test muy bueno.

[0.97, 1): Test excelente.

Regresión Logística

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i = z_i$$

$$p_i = \frac{e^{z_i}}{(1 + e^{z_i})}$$

Razón de chances: Por ejemplo, si la probabilidad de que ocurra un evento es del 80% entonces se espera que haya 4 veces más probabilidad de que ocurra que de que NO ocurra

$$\frac{0.8}{0.2} = 4$$

Regresion Probit

$$\hat{y}_i = \Phi(\beta_0 + \beta_1 x_i)^{-1}$$