CENTRO UNIVERSITARIO
DE TECNOLOGÍA
Y ARTE DIGITAL

Tema 3

Estadística Descriptiva II
Distribuciones Bidimensionales

Mar Angulo Martínez
mar.angulo@u-tad.com

CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. **Variables bidimensionales.**
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Variables bidimensionales

Variable estadística bidimensional:

- Surge cuando se estudian 2 características en una misma población.
- Es decir, dos variables estadísticas (X, Y), donde
    - X toma los valores $\{x_1, x_2, \dots, x_n\}$ y
    - Y toma los valores $\{y_1, y_2, \dots, y_n\}$
- Así, las observaciones de la variable estadística bidimensional serán $(x_i,\ y_i)$,
    - donde $i = \{1, \dots, n\}$
    - siendo n el número de individuos muestreados (el tamaño muestral).
- X e Y pueden ser discretas o continuas o una discreta y una continua
- La variable estadística bidimensional es cuantitativa cuando X e Y toman valores numéricos

Ejemplos:

Muestra de 500 alumnos → (X = altura (cm), Y = talla de zapato)
Muestra de 50 familias → (X = salario total, Y = número de hijos)
Muestra de 1000 empleados → (X = años trabajados en la empresa, Y = remuneración anual)
Muestra de 30 niños de una escuela → (X = peso en kg, Y = altura en cm)

U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. **Tablas de contingencia.**
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Tablas de contingencia

- A partir de una serie de observaciones $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, se pueden agrupar los datos de las frecuencias para mostrarlos en una tabla bidimensional en k y m grupos, respectivamente, llamada **tabla de contingencia**.

- Para el caso discreto,

$X \mid Y$	y_1	y_2	…	y_m
x_1	n_{11}	n_{12}	$\dots$	n_{1m}
x_2	n_{21}	n_{22}	$\dots$	n_{2m}
…	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{km}

De la misma forma que creamos una tabla de frecuencias absolutas conjunta, también podemos crear una tabla de frecuencias relativas conjunta f_{ij}

- donde n_{ij} representa la frecuencia absoluta de observaciones (x_i, y_j),
  - con $i = \{1, \dots, k\}$ y $j = \{1, \dots, m\}$.

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. **Distribuciones marginales y condicionadas.**
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Distribuciones marginales

Ejemplo: Dada la siguiente tabla de frecuencias, creada a partir de la encuesta a un conjunto de personas, relacionando X ="edad del encuestado" e Y ="número de monedas que lleva en el bolsillo",

$X \mid Y$	0	2	4	6	8
16	25	5	3	0	0
18	10	12	16	5	1
20	1	2	5	1	5
25	0	0	5	19	16

¿Cuántas personas de las encuestadas tenían 18 años?
¿Cuántas personas de las encuestadas tenían 4 monedas en el bolsillo?

Distribuciones marginales

Ejemplo: Dada la siguiente tabla de frecuencias, creada a partir de la encuesta a un conjunto de personas, relacionando X =“edad del encuestado” e Y =“número de monedas que lleva en el bolsillo”,

¿Cuántas personas de las encuestadas tenían 18 años?

Es una pregunta relacionada con la variable X, pero para su respuesta no se ha necesitado conocimiento alguno sobre la variable Y.

¿Cuántas personas de las encuestadas tenían 4 monedas en el bolsillo?

Análogamente, es una pregunta relacionada con la variable Y, pero para su respuesta no se ha necesitado conocimiento alguno sobre la variable X.

U-tad
CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Distribuciones marginales

Ejemplo: Dada la siguiente tabla de frecuencias, creada a partir de la encuesta a un conjunto de personas, relacionando X =“edad del encuestado” e Y =“número de monedas que lleva en el bolsillo”,

- Concretamente, la distribución marginal de X es una variable estadística unidimensional que tiene, en este caso, la siguiente tabla de frecuencias:

x_i		$n_{x_i} = n_{i.}$
16	$25 + 5 + 3 + 0 + 0$	33
18	$10 + 12 + 16 + 5 + 1$	44
20	$1+2+5+1+5$	14
25	$0 + 0 + 5 + 19 + 16$	40

U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Distribuciones marginales

Ejemplo: Dada la siguiente tabla de frecuencias, creada a partir de la encuesta a un conjunto de personas, relacionando X ="edad del encuestado" e Y ="número de monedas que lleva en el bolsillo",

$X \mid Y$	0	2	4	6	8
16	25	5	3	0	0
18	10	12	16	5	1
20	1	2	5	1	5
25	0	0	5	19	16

- Por tanto, las distribuciones marginales de este ejercicio son:

X	$n_{x_i} = n_{i.}$
16	33
18	44
20	14
25	40

Y	$n_{y_j} = n_{.j}$
0	36
2	19
4	29
6	25
8	22

U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Distribuciones condicionadas

Ejemplo: Dada la siguiente tabla de frecuencias, creada a partir de la encuesta a un conjunto de personas, relacionando X ="edad del encuestado" e Y ="número de monedas que lleva en el bolsillo",

$X \mid Y$	0	2	4	6	8
16	25	5	3	0	0
18	10	12	16	5	1
20	1	2	5	1	5
25	0	0	5	19	16

¿Cuál es la distribución de monedas en las personas encuestadas de 18 años? ¿Cómo es la población, en función de la edad, de personas que llevan 4 monedas en el bolsillo?

Distribuciones condicionadas

Ejemplo: Dada la siguiente tabla de frecuencias, creada a partir de la encuesta a un conjunto de personas, relacionando X ="edad del encuestado" e Y ="número de monedas que lleva en el bolsillo",

$X \mid Y$	0	2	4	6	8
16	25	5	3	0	0
18	10	12	16	5	1
20	1	2	5	1	5
25	0	0	5	19	16

$(Y|X = 18)$

$(X|Y = 4)$

- Lo que observamos es que cada una de estas condiciones tiene su propia distribución de frecuencias:
  - $X|Y = y_j$ → distribución de X sabiendo que Y vale y_j
  - $Y|X = x_i$ → distribución de Y sabiendo que X vale x_i

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. **Covarianza.**
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Covarianza

La **Covarianza muestral de una variable aleatoria bidimensional (s_{xy})** mide el grado de interdependencia o asociación entre las 2 variables mediante cuantificar la variabilidad conjunta entre X e Y.

$$s_{xy} = Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} x_i\, y_j - \bar{x}\bar{y}$$

- Si $s_{xy} > 0$ → *dependencia directa (positiva)*
  - *a grandes valores de x, grandes valores de y.*
- *Si* $s_{xy} = 0$ → *independientes*
  - *no existe una relación lineal entre las dos variables estudiadas.*
- *Si* $s_{xy} < 0$ → *dependencia inversa (negativa)*
  - *a grandes valores de x, pequeños valores de y.*

U-tad CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Temario

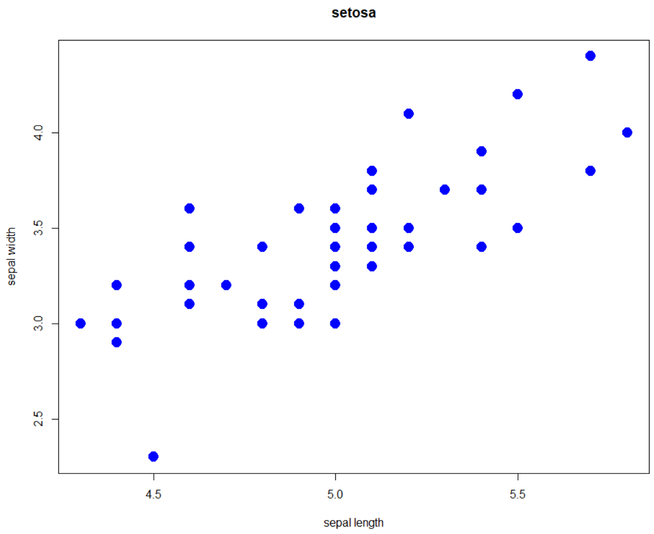
TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. **Visualización de los datos. Gráficas de datos bidimensionales.**
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Gráficas estadísticas (visualización de datos)

Nube de puntos (*scatterplot*):

SOLO variables numéricas



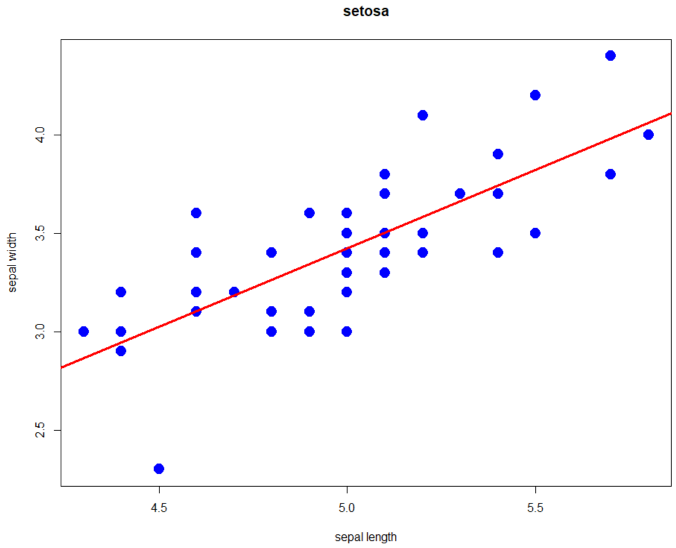

```r
# Lectura del dataframe
datos = iris
head(datos)
dim(datos)

# Filtrado de datos
filter.datos = datos[datos$Species == 'setosa',]
dim(filter.datos)

# Scatterplot
x = filter.datos$Sepal.Length
y = filter.datos$Sepal.Width

plot(x, y,
     col = 'blue',
     pch = 19,
     cex = 2,
     xlab = "sepal length", ylab = "sepal width",
     main = "setosa")
```

¿Existe alguna relación entre X e Y?

Gráficas estadísticas (visualización de datos)

Recta de regresión:

SOLO variables numéricas




```r
# calculo de la recta de regresion
regresion = lm(formula = y ~ x)
regresion

plot(x, y,
     col = 'blue',
     pch = 19,
     cex = 2,
     xlab = "sepal length", ylab = "sepal width",
     main = "setosa")

# recta de regresión (sobre scatterplot)
abline(regresion,
       col = 'red',
       lwd = 3)
```

Cuanto mayor es X, mayor es Y (y viceversa) → existe una relación entre variables

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. **El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.**
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

Regresión lineal

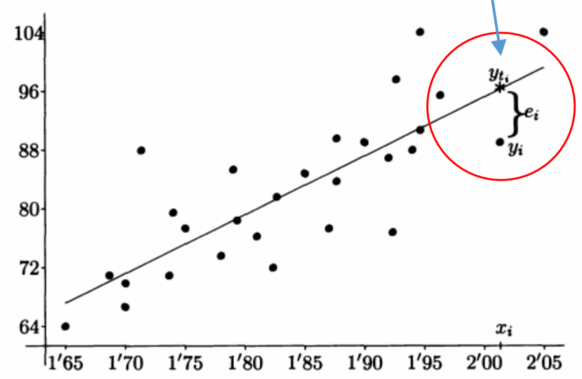
La **recta de regresión lineal** es la recta que mejor se aproxima a la nube de puntos de la muestra.

$$r: \quad y = \beta_1 x + \beta_0$$

valor estimado de y_i para x_i

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

Para cada punto de la nube de puntos (x_i, y_i), **el residuo** e_i es la distancia entre dicho punto y el valor estimado por la recta de regresión $(x_i, \beta_1 x_i + \beta_0)$.



Regresión lineal (mínimos cuadrados)

El cálculo de la recta de regresión lineal se realiza con el **método de los Mínimos cuadrados**, que es el siguiente:

- Se trata de encontrar aquella recta tal que

$$e_i = 0, \forall i$$

- Dado que no siempre es posible, se trata al menos de resolver el siguiente problema de optimización

$$min \sum_{i=1}^{n} {e_i}^2$$

- Por tanto, para calcular la recta de regresión, el problema se reduce a calcular los valores $\beta_0, \beta_1 \in \mathbb{R}$, de la recta de regresión, tales que minimicen la suma de los residuos.

Regresión lineal (mínimos cuadrados)

- Por tanto, para resolver la siguiente ecuación:

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} {e_i}^2 = \min_{\beta_0,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} (\widehat{y_i} - y_i)^2 = \min_{\beta_0,\beta_1\in\mathbb{R}} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2$$

- Derivamos respecto de las variables que tenemos que calcular: β_0, β_1.

$$0 = \frac{\partial \sum_{i=1}^{n}(\beta_1 x_i + \beta_0 - y_i)^2}{\partial \beta_0} \ [1] \quad \text{y} \quad 0 = \frac{\partial \sum_{i=1}^{n}(\beta_1 x_i + \beta_0 - y_i)^2}{\partial \beta_1} \ [2]$$

U-tad
CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Regresión lineal (mínimos cuadrados)

- En [1]:

$$\frac{\partial(\beta_1 x + \beta_0 \ - y)^2}{\partial \beta_0} = 2(\beta_1 x + \beta_0 \ - y) = 0$$

- Por tanto, [1] quedaría así:

$$0 = \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 \ - y_i)$$

- Despejamos β_0 :

$$0 = \sum_{i=1}^{n} \beta_1 x_i + \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} y_i$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i$$

$$\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

Regresión lineal (mínimos cuadrados)

- En [2]:

$$\frac{\partial(\beta_1 x + \beta_0 \ - y)^2}{\partial \beta_1} = 2(\beta_1 x + \beta_0 \ - y)x = 0$$

- Por tanto, [2] quedaría así:

$$0 = \sum_{i=1}^{n} x_i(\beta_1 x_i + \beta_0 \ - y_i) = \sum_{i=1}^{n} (\beta_1 {x_i}^2 + \beta_0 x_i - y_i x_i)$$

Regresión lineal (mínimos cuadrados)

- Sustituimos β_0 de [1] en [2]:

$$0 = \beta_1 \sum_{i=1}^{n} {x_i}^2 + \beta_0 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\, y_i$$

$$0 = \beta_1 \sum_{i=1}^{n} {x_i}^2 + n\beta_0 \bar{x} - \sum_{i=1}^{n} x_i\, y_i$$

$$0 = \beta_1 \sum_{i=1}^{n} {x_i}^2 + n(\bar{y} - \beta_1 \bar{x})\bar{x} - \sum_{i=1}^{n} x_i\, y_i$$

Regresión lineal (mínimos cuadrados)

- Despejamos β_1:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n}\, x_i{}^2\, - n\bar{x}^2}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{y}\bar{x}}{n s_x^2}$$

$$\boxed{\beta_1 = \frac{s_{xy}}{s_x^2}}$$

Regresión lineal

La **recta de regresión de $\boldsymbol{Y}$ sobre $\boldsymbol{X}$** se calcula, por tanto,

$$\boxed{y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})}$$

De forma análoga con el método de los mínimos cuadrados, podría calcularse la **recta de regresión de $\boldsymbol{X}$ sobre $\boldsymbol{Y}$**,

$$\boxed{x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})}$$

Regresión lineal

El **coeficiente de regresión de $\boldsymbol{Y}$ sobre $\boldsymbol{X}$**

$$\boxed{b_{yx} = \frac{s_{xy}}{s_x^2}}$$

Es la pendiente de la **recta de regresión de $\boldsymbol{Y}$ sobre $\boldsymbol{X}$**

- Representa el incremento de y por cada aumento unitario de x

Regresión lineal

El **coeficiente de regresión de X sobre Y**

$$\boxed{b_{xy} = \frac{s_{xy}}{s_y^2}}$$

De forma análoga con el método de los mínimos cuadrados, podría calcularse la **recta de regresión de X sobre Y**,

Regresión lineal

La **recta de regresión de Y sobre X** se calcula, por tanto,

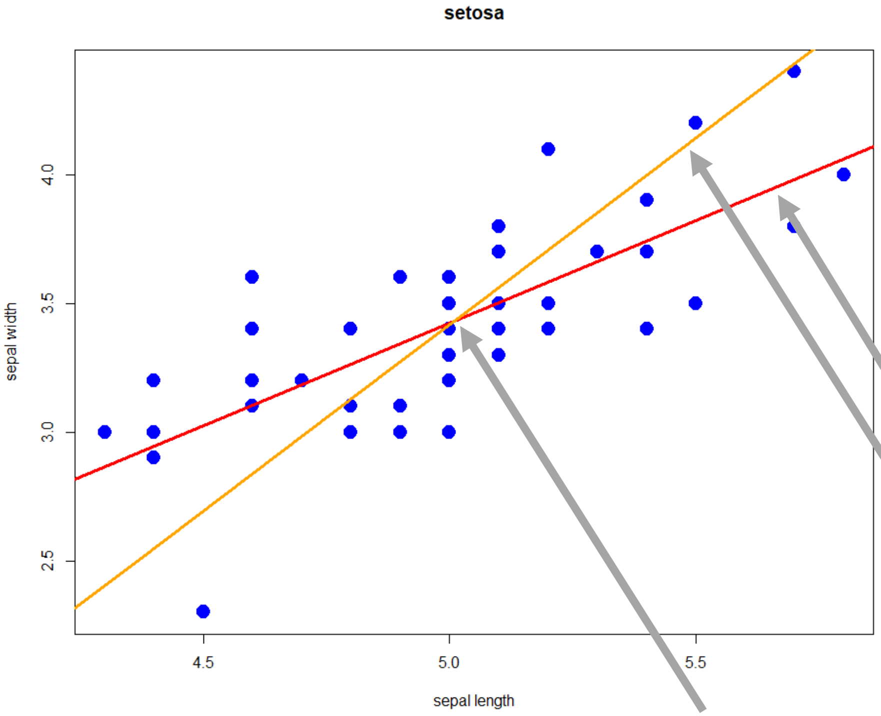
$$\boxed{y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})}$$

De forma análoga con el método de los mínimos cuadrados, podría calcularse la **recta de regresión de X sobre Y**,

$$\boxed{x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})}$$

U-TAD
CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Regresión lineal




```r
# Scatterplot
x = filter.datos$Sepal.Length
y = filter.datos$Sepal.Width

plot(x, y,
     col = 'blue',
     pch = 19,
     cex = 2,
     xlab = "sepal length", ylab = "sepal width",
     main = "setosa")

# cálculo de la recta de regresion
regresion1 = lm(formula = y ~ x)
regresion2 = lm(formula = x ~ y)

plot(x, y,
     col = 'blue',
     pch = 19,
     cex = 2,
     xlab = "sepal length", ylab = "sepal width",
     main = "setosa")

# recta de regresión (y sobre x)
abline(regresion1,
       col = 'red',
       lwd = 3)

# recta de regresión (x sobre y)
y = seq(2, 7, 0.25)
x = regresion2$coefficients[1] + regresion2$coefficients[2] * y
lines(x, y,
       col = 'orange',
       lwd = 3)
```

centro de gravedad

Regresión lineal

Propiedades de las rectas de regresión

- $b_{yx} = \frac{s_{xy}}{s_x^2}$ es el coeficiente de regresión de y sobre x y es el incremento que experimenta la variable y cuando la variable x aumenta en una unidad
- $b_{xy} = \frac{s_{xy}}{s_y^2}$ es el coeficiente de regresión de x sobre y :es el incremento que experimenta la variable x cuando la variable y aumenta en una unidad
- $\frac{s_{xy}}{s_x^2}$ y $\frac{s_{xy}}{s_y^2}$ son las pendientes de las rectas de regresión.

  - Si $s_{xy} > 0$ → *dependencia directa (positiva)*
    - *a grandes valores de x, grandes valores de y.*
  - *Si* $s_{xy} = 0$ → *independientes*
    - *no existe una relación lineal entre las dos variables estudiadas.*
  - *Si* $s_{xy} < 0$ → *dependencia inversa (negativa)*
    - *a grandes valores de x, pequeños valores de y.*

- Las dos rectas se cortan en el llamado centro de gravedad de la distribución que coincide con el punto $(\bar{x}, \bar{y})$.

U-tad
CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Regresión lineal

Ejemplo 1:

Dados los siguientes datos sobre ingresos y consumo de un grupo de familias al cabo de un mes (en cientos de euros), se pide construir las rectas de regresión:

Ingresos	8	15	20	25	25	25	8	13	7	6	12	15
Consumo	10	14	13	5	10	20	8	5	12	8	10	14

Nuestro objetivo será,

1. Calcular los estadísticos necesarios → medias, varianzas, covarianzas
2. Calcular las rectas

Regresión lineal

Ejemplo 1:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
8	10	64	100	80
15	14	225	196	210
20	13	400	169	260
25	5	625	25	125
25	10	625	100	250
25	20	625	400	500
8	8	64	64	64
13	5	169	25	65
7	12	49	144	84
6	8	36	64	48
12	10	144	100	120
15	14	225	196	210
179	**129**	**3.251**	**1.583**	**2.016**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{179}{12} = 14{,}92$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{129}{12} = 10{,}75$$

$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 = \frac{3251}{12} - (14{,}917)^2 = 48{,}4$$

$$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2 = \frac{1583}{12} - (10{,}75)^2 = 16{,}35$$

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = \frac{2016}{12} - 14{,}92 \cdot 10{,}75 = 7{,}64$$

Regresión lineal

Ejemplo 1:

- Recta de regresión de Y sobre X :

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \Rightarrow y\text{−}10{,}75\text{=}\frac{7{,}64}{48{,}4}\ (x - 14{,}92) \Rightarrow y = 0{,}158x + 8{,}392$$

➢ Por cada 100 euros más de ingresos, estas familias gastan 15,79 euros más

- Recta de regresión de X sobre Y:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \Rightarrow x\text{−}14{,}92\text{=}\frac{7{,}64}{16{,}35}\ (x - 10{,}75) \Rightarrow x = 0{,}467y + 9{,}9$$

➢ Por cada 100 euros más de consumo, los ingresos familiares son 46,7 euros mayores

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. **Correlación lineal.**
8. Otros modelos de regresión y correlación.
9. Bondad de ajuste.

U-tad
CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

Correlación lineal

El **Coeficiente de correlación lineal de Pearson** $(\boldsymbol{r_{xy}} = \boldsymbol{r})$ mide el grado de interdependencia entre las dos variables X e Y

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Nota:

- Si $r > 0$ → dependencia directa (positiva)
- Si $r = 0$ → independientes
- Si $r < 0$ → dependencia inversa (negativa)

Importante:

$$\boxed{-1 \leq r \leq 1}$$

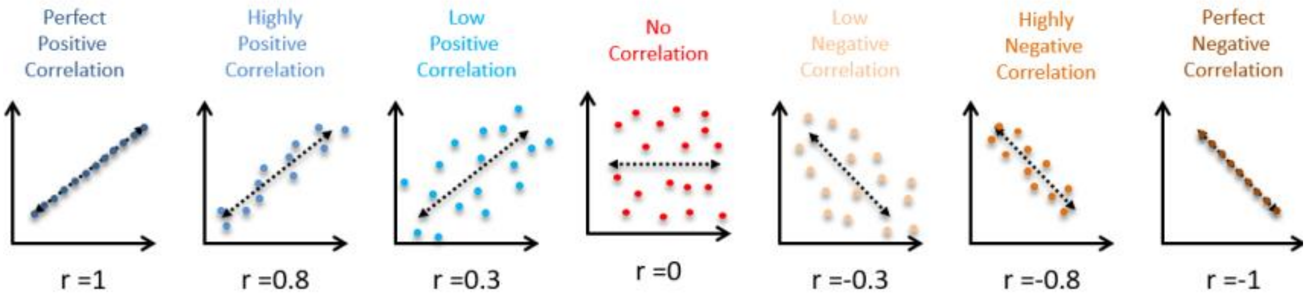
Correlación lineal (interpretación)

Caso 1: r = +1 ó -1

- Interdependencia total y directa (r=1) o inversa (r=-1) entre las variables
- Los valores observados coinciden con los teóricos y la recta de regresión pasa por los puntos
- Toda la variabilidad de Y queda explicada por X mediante el ajuste
- El ajuste es perfecto, todos los errores son 0 y la fiabilidad de las predicciones será del 100%
- Las 2 rectas de regresión son coincidentes

Caso 2: r = 0

- Interdependencia nula: variables incorreladas. La independencia lineal es nula
- Los valores observados están completamente alejados de los teóricos y la recta de regresión pasa muy alejada de la nube de puntos
- El ajuste es nulo y la fiabilidad de cualquier predicción que hagamos con él será 0
- Ninguna variabilidad de Y queda explicada por X mediante el ajuste
- Las 2 rectas de regresión son perpendiculares



Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. **Otros modelos de regresión y correlación.**
9. Bondad de ajuste.

Regresión múltiple y no lineal

Regresión lineal múltiple:

- Es una regresión que incluye más de una variable independiente

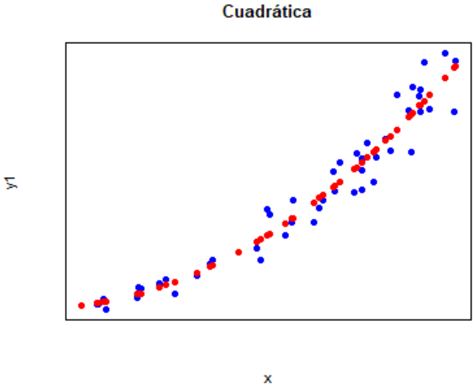
$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

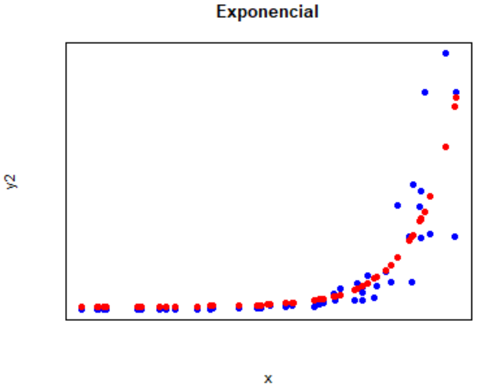
Regresión múltiple y no lineal

Regresión no lineal:

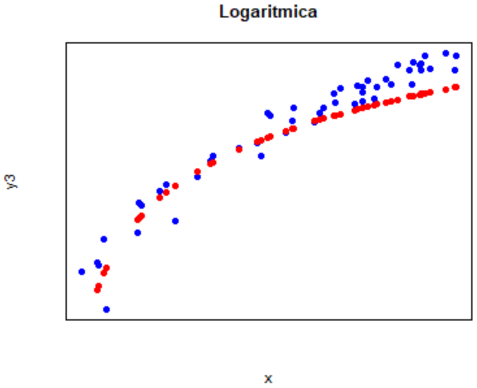
- Es una regresión ajusta la nube de puntos por una función no lineal



$$y = \beta_2 x^2 + \beta_1 x + \beta_0$$



$$y = e^{\beta_1 x + \beta_0}$$



$$y = log(\beta_1 x + \beta_0)$$

Temario

TEMA 3: Estadística descriptiva 2. Distribuciones bidimensionales.

1. Variables bidimensionales.
2. Tablas de contingencia.
3. Distribuciones marginales y condicionadas.
4. Covarianza.
5. Visualización de los datos. Gráficas de datos bidimensionales.
6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados.
7. Correlación lineal.
8. Otros modelos de regresión y correlación.
9. **Bondad de ajuste.**

Bondad de ajuste

- Así, la nube de puntos no siempre se aproxima a una recta.
- Para saber si la nube de puntos se rige por un comportamiento lineal, cuadrático, exponencial o logarítmico, **necesitamos una medida que explique lo próxima que está la nube de puntos a la curva estimada**.
- Estas medidas serán siempre funciones de los errores e_i.

Bondad de ajuste:

- Es el criterio que nos permite medir describir el ajuste entre un conjunto de observaciones y una curva de regresión.
- Existen varios estadísticos para medir estos criterios. En regresión, el más utilizado es el **coeficiente de determinación**.

Bondad de ajuste

El **Coeficiente de determinación lineal** $(\boldsymbol{R^2})$ mide el grado de fiabilidad que hagamos con cualquier ajuste de regresión.

$$R^2 = 1 - \frac{V_r}{s_y^2} \leftrightsquigarrow R^2 = 1 - \frac{V_r}{s_x^2}$$

Donde, $V_r = \frac{1}{n}\sum_{i=1}^{n} {e_i}^2$ es la varianza de los errores cometidos. Por tanto, en una recta de regresión de Y sobre X, representa el porcentaje de variabilidad de los errores cometidos en el ajuste sobre la varianza marginal de Y.

En el caso de la regresión lineal simple, $\boldsymbol{R^2 = r^2}$

- R^2 siempre es positivo.
- Cuanto más se acerque al valor 1, mejor será el ajuste.
- Cuanto más se acerque al valor 0, peor será el ajuste.

Bondad de ajuste

Ejemplo 1:

Datos sobre ingresos y consumo de un grupo de familias al cabo de un mes (en cientos de euros)

Interdependencia entre las variables:

$$r = \frac{s_{xy}}{s_x . s_y} = \frac{7{,}64}{6{,}96 \cdot 4{,}044} = +0{,}27$$

- Interdependencia directa (a mayores ingresos, mayor consumo, y viceversa)
- Interdependencia más bien débil entre X e Y

Bondad de ajuste. Fiabilidad de las predicciones:

$$R^2 = 0{,}0729$$

- Las predicciones que hagamos con este ajuste tendrán una fiabilidad del 7,29%
- Valores teóricos muy alejados de los observados y recta de regresión que pasa muy alejada de los puntos de la nube

CENTRO UNIVERSITARIO DE TECNOLOGÍA Y ARTE DIGITAL

Regresión (notas finales)

La regresión es un modelo matemático utilizado principalmente para:

- **Explicar** la relación natural entre dos (o más) variables de una población.
- **Predecir** el comportamiento de una variable en función del conocimiento de la otra.
- Las variables conocidas (x_i) se llaman **variables independientes**.
- La variable desconocida (y) se llama **variable dependiente**.

El tipo de regresión depende de:

- La curva elegida: Lineal (recta, plano, hiperplano), Cuadrática (polinomio grado 2), Polinomial (polinomio de grado superior), Exponencial, Logarítmica, etc.
- La cantidad de variables dependientes: Una variable (simple) o varias variables (múltiple).