TEMA 2: Estadística descriptiva bidimensional

- Distribución conjunta de dos variables estadísticas. Tablas y representaciones gráficas.
- Distribuciones marginales.
- Distribuciones condicionadas.
- Dependencia funcional. Dependencia e independencia estadística.
- Momentos bidimensionales.
- Regresión y correlación:
    - Ajuste de funciones por mínimos cuadrados. Curvas y rectas de regresión.
    - Razón de correlación. Correlación lineal.

DISTRIBUCIÓN CONJUNTA DE DOS CARACTERES

Consideramos una población formada por n individuos en la que se estudian simultáneamente dos variables (atributos), X e Y, con valores (modalidades) $x_1, \cdots, x_k$ e $y_1, \ldots, y_p$, respectivamente.

- **Frecuencia absoluta del par (x_i, y_j):**

  n_{ij} = número de individuos de la población que presentan simultáneamente los valores (modalidades) x_i de X e y_j de Y.

  Distribución de frecuencias absolutas: $\Big\{((x_i, y_j),\ n_{ij});\ i = 1, \ldots, k,\ j = 1, \ldots, p\Big\}$

$$\sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} = n.$$

- **Frecuencia relativa del par (x_i, y_j):**

  f_{ij} = proporción de individuos de la población que presentan simultáneamente los valores (modalidades) x_i de X e y_j de Y $\longrightarrow$ $f_{ij} = \dfrac{n_{ij}}{n}$.

  Distribución de frecuencias relativas: $\Big\{((x_i, y_j),\ f_{ij});\ i = 1, \ldots, k,\ j = 1, \ldots, p\Big\}$

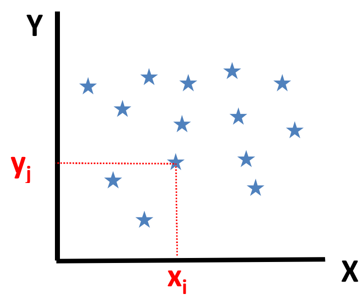
$$\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} = 1.$$

Tabla de frecuencias bidimensionales

$X \quad \backslash \ Y$	y_1	$\cdots$	y_j	$\cdots$	y_p
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1p}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ip}
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kp}

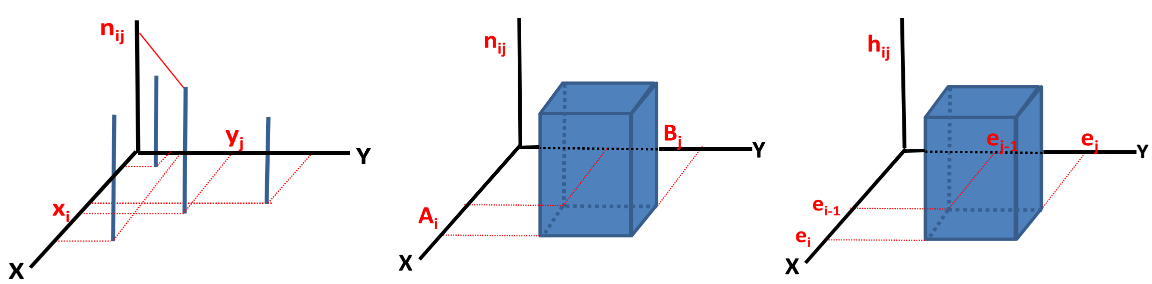
REPRESENTACIONES GRÁFICAS

- **Diagrama de dispersión o nube de puntos** (*datos cuantitativos*): Se representa cada par de observaciones (o marcas de clase, si los datos están agrupados en intervalos) (x_i, y_j) en un sistema de ejes cartesianos. Si un par tiene frecuencia mayor que uno, ésta puede indicarse al lado del punto correspondiente:



- **Estereograma**: gráfico tridimensional formado por barras o prismas según los casos:
  - *Variables discretas*: barras verticales de longitudes iguales (o proporcionales) a las frecuencias de cada par de valores, que se representan en el plano base.
  - *Caracteres cualitativos*: prismas con base común y alturas iguales (o proporcionales) a las frecuencias de cada par de modalidades, representadas en el plano base.
  - *Variables continuas*: prismas con bases definidas por cada par de intervalos y alturas iguales (o proporcionales) a las *densidades de frecuencia*, de forma que el volumen de cada prisma es igual (o proporcional) a la frecuencia de la pareja de intervalos correspondiente.

$$h_{ij} = \frac{n_{ij}}{(e_i^x - e_{i-1}^x)(e_j^y - e_{j-1}^y)}$$



DISTRIBUCIONES MARGINALES

$$\Big\{((x_i, y_j),\ n_{ij});\ i = 1, \ldots, k,\ j = 1, \ldots, p\Big\}, \quad n = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}, \quad f_{ij} = \frac{n_{ij}}{n}$$

Frecuencia absoluta del valor (modalidad) x_i del carácter X: número de individuos que presentan dicho valor (sin tener en cuenta el que presentan en Y):

$$n_{i.} = \sum_{j=1}^{p} n_{ij}\ ;\quad i = 1, 2, \ldots, k \quad \longrightarrow \quad \sum_{i=1}^{k} n_{i.} = n$$

Frecuencia relativa del valor (modalidad) x_i del carácter X: proporción de individuos que presentan dicho valor (sin tener en cuenta el que presentan en Y):

$$f_{i.} = \frac{n_{i.}}{n} = \sum_{j=1}^{p} f_{ij}\ ;\quad i = 1, 2, \ldots, k \quad \longrightarrow \quad \sum_{i=1}^{k} f_{i.} = 1$$

Distribución marginal de X: $\{(x_i, n_{i.}),\ i = 1, 2, \ldots, k\} \ \equiv\ \{(x_i, f_{i.}),\ i = 1, 2, \ldots, k\}$

Frecuencia absoluta del valor (modalidad) y_j del carácter Y: número de individuos que presentan dicho valor (sin tener en cuenta el que presentan en X):

$$n_{.j} = \sum_{i=1}^{k} n_{ij}\ ;\quad j = 1, 2, \ldots, p \quad \longrightarrow \quad \sum_{j=1}^{p} n_{.j} = n$$

Frecuencia relativa del valor (modalidad) y_j del carácter Y: proporción de individuos que presentan dicho valor (sin tener en cuenta el que presentan en X):

$$f_{.j} = \frac{n_{.j}}{n} = \sum_{i=1}^{k} f_{ij}\ ;\quad j = 1, 2, \ldots, p \quad \longrightarrow \quad \sum_{j=1}^{p} f_{.j} = 1$$

Distribución marginal de Y: $\{(y_j, n_{.j}),\ j = 1, 2, \ldots, p\} \ \equiv\ \{(y_j, f_{.j}),\ j = 1, 2, \ldots, p\}$

$X \ \backslash\ Y$	y_1	$\cdots$	y_j	$\cdots$	y_p	*Sumas*
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$\mathbf{n_{1.}}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$\mathbf{n_{i.}}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$\mathbf{n_{k.}}$
Sumas	$\mathbf{n_{.1}}$	$\cdots$	$\mathbf{n_{.j}}$	$\cdots$	$\mathbf{n_{.p}}$	n

$$\overline{x} = \sum_{i=1}^{k} f_{i.} x_i, \qquad \sigma_x^2 = \sum_{i=1}^{k} f_{i.}(x_i - \overline{x})^2$$

$$\overline{y} = \sum_{j=1}^{p} f_{.j} y_j, \qquad \sigma_y^2 = \sum_{j=1}^{p} f_{.j}(y_j - \overline{y})^2$$

DISTRIBUCIONES CONDICIONADAS

$X \quad \backslash \; Y$	y_1	$\cdots$	y_j	$\cdots$	y_p	$n_{i.}$
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	n

Distribución condicionada de X a $Y = y_j; \quad j = 1, \dots p$

$\downarrow$

Sólo se consideran los $n_{.j}$ individuos que presentan el valor (o modalidad) y_j en el carácter Y (columna j-ésima).

- Frecuencia absoluta de x_i en los individuos tales que $Y = y_j \longrightarrow n_{ij}, \quad i = 1, \dots k$.
- Frecuencia relativa de x_i en los individuos tales que $Y = y_j \longrightarrow f_{i/j} = \dfrac{n_{ij}}{n_{.j}}, \quad i = 1, \dots k$.

$X/Y = y_j$	n_{ij}	$f_{i/j}$
x_1	n_{1j}	$f_{1/j}$
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	$f_{i/j}$
$\vdots$	$\vdots$	$\vdots$
x_k	n_{kj}	$f_{k/j}$
Sumas	$n_{.j}$	1

- $f_{i/j} = \dfrac{n_{ij}}{n_{.j}} = \dfrac{f_{ij}}{f_{.j}} \longrightarrow f_{ij} = f_{i/j} f_{.j}, \; i = 1, \dots, k, \;\; j = 1, \dots, p.$

Media condicionada: $\overline{x}_j = \sum_{i=1}^{k} f_{i/j} x_i$

Varianza condicionada: $\sigma^2_{x,j} = \sum_{i=1}^{k} f_{i/j} (x_i - \overline{x}_j)^2$

$$\boxed{\overline{x} = \sum_{j=1}^{p} f_{.j} \overline{x}_j, \qquad \sigma^2_x = \sum_{j=1}^{p} f_{.j} \sigma^2_{x,j} + \sum_{j=1}^{p} f_{.j} (\overline{x}_j - \overline{x})^2}$$

Distribución condicionada de Y a $X = x_i; \quad i = 1, \dots k$

$\downarrow$

Sólo se consideran los $n_{i.}$ individuos que presentan el valor (o modalidad) x_i en el carácter X (fila i-ésima).

- Frecuencia absoluta de y_j en los individuos tales que $X = x_i \longrightarrow n_{ij}, \quad j = 1, \dots p$
- Frecuencia relativa de y_j en los individuos tales que $X = x_i \longrightarrow f_{j/i} = \dfrac{n_{ij}}{n_{i.}}, \quad j = 1, \dots p$

$Y/X = x_i$	n_{ij}	$f_{j/i}$
y_1	n_{i1}	$f_{1/i}$
$\vdots$	$\vdots$	$\vdots$
x_j	n_{ij}	$f_{j/i}$
$\vdots$	$\vdots$	$\vdots$
y_p	n_{ip}	$f_{p/i}$
Sumas	$n_{i.}$	1

- $f_{j/i} = \dfrac{n_{ij}}{n_{i.}} = \dfrac{f_{ij}}{f_{i.}} \longrightarrow f_{ij} = f_{j/i} f_{i.}, \; i = 1, \dots, k, \;\; j = 1, \dots, p.$

Media condicionada: $\overline{y}_i = \sum_{j=1}^{p} f_{j/i} y_j$

Varianza condicionada: $\sigma^2_{y,i} = \sum_{j=1}^{p} f_{j/i} (y_j - \overline{y}_i)^2$

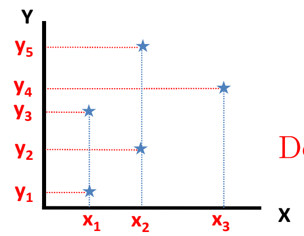
$$\boxed{\overline{y} = \sum_{i=1}^{k} f_{i.} \overline{y}_i, \qquad \sigma^2_y = \sum_{i=1}^{k} f_{i.} \sigma^2_{y,i} + \sum_{i=1}^{k} f_{i.} (\overline{y}_i - \overline{y})^2}$$

DEPENDENCIA FUNCIONAL Y DEPENDENCIA ESTADÍSTICA

DEPENDENCIA FUNCIONAL

Un carácter X depende funcionalmente de un carácter Y si la modalidad que presenta cada individuo en Y determina la que presenta en X (para cada modalidad de Y sólo hay una modalidad de X con frecuencia no nula). Para caracteres cuantitativos, $X = f(Y)$:

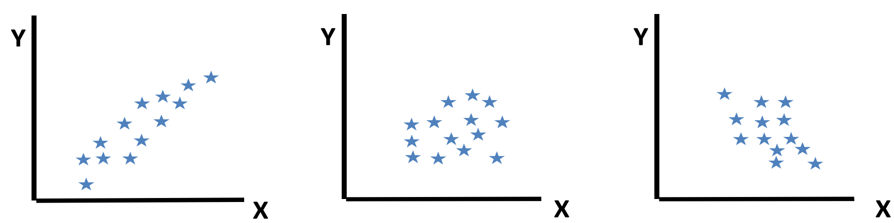
$X \quad \backslash \; Y$	y_1	y_2	y_3	y_4	y_5
x_1	n_{11}	0	n_{13}	0	0
x_2	0	n_{22}	0	0	n_{25}
x_3	0	0	0	n_{34}	0



Dependencia no recíproca

DEPENDENCIA ESTADÍSTICA

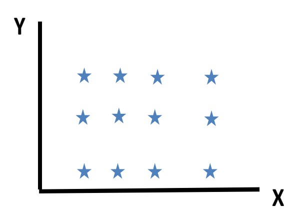
Un carácter X es estadísticamente dependiente de un carácter Y si el comportamiento de X es diferente según las modalidades de Y:



Dependencia recíproca

$$\text{DEPENDENCIA FUNCIONAL} \;\; \begin{array}{c} \Longrightarrow \\ \not\Longleftarrow \end{array} \;\; \text{DEPENDENCIA ESTADÍSTICA}$$

Un carácter X es **estadísticamente independiente** *de un carácter Y si el comportamiento de X es el mismo para todas las modalidades de Y* (propiedad recíproca). En tal caso, el diagrama de dispersión es un rectángulo (condición necesaria para la independencia):



INDEPENDENCIA ESTADÍSTICA

$$\Big\{((x_i, y_j), n_{ij});\ i = 1, \ldots, k,\ j = 1, \ldots, p\Big\} \quad n = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}, \quad f_{ij} = \frac{n_{ij}}{n}$$

- $n_{i.} = \sum_{j=1}^{p} n_{ij},\ i = 1, \ldots, k; \quad n_{.j} = \sum_{i=1}^{k} n_{ij},\ j = 1, \ldots, p$
- $f_{i.} = \dfrac{n_{i.}}{n},\ i = 1, \ldots, k; \quad f_{.j} = \dfrac{n_{.j}}{n},\ j = 1, \ldots, p$
- $f_{i/j} = \dfrac{n_{ij}}{n_{.j}} = \dfrac{f_{ij}}{f_{.j}}; \quad f_{j/i} = \dfrac{n_{ij}}{n_{i.}} = \dfrac{f_{ij}}{f_{i.}}, \quad i = 1, \ldots, k,\ j = 1, \ldots, p$

- *El carácter X es (estadísticamente) independiente del carácter Y si las distribuciones de X condicionadas a cada modalidad de Y coinciden:*

$$\forall i = 1, \ldots, k, \quad f_{i/j} \text{ no depende de } j \quad (f_{i/1} = f_{i/2} = \cdots = f_{i/p}, \quad \forall i = 1, \ldots, k).$$

- *Similarmente, Y es (estadísticamente) independiente de X si las distribuciones de Y condicionadas a cada modalidad de X coinciden:*

$$\forall j = 1, \ldots, p, \quad f_{j/i} \text{ no depende de } i \quad (f_{j/1} = f_{j/2} = \cdots = f_{j/k}, \quad \forall j = 1, \ldots, p).$$

Caracterizaciones de independencia:

- *X es independiente de Y si y sólo si la distribución condicionada de X a cualquier modalidad de Y coincide con la distribución marginal de X:*

$$X \text{ es independiente de } Y \Leftrightarrow f_{i/j} = f_{i.}, \quad \forall i = 1, \ldots, k, \quad \forall j = 1, \ldots, p.$$

- *X es independiente de Y si y sólo si Y es independiente de X.*

- *X e Y son independientes* $\Leftrightarrow\ f_{ij} = f_{i.}f_{.j},\ \forall i = 1, \ldots, k, \quad \forall j = 1, \ldots, p$

$$\Leftrightarrow\ n_{ij} = \frac{n_{i.}n_{.j}}{n},\ \forall i = 1, \ldots, k, \quad \forall j = 1, \ldots, p.$$

MOMENTOS BIDIMENSIONALES

$$\Big\{((x_i, y_j), n_{ij});\ i = 1, \dots, k,\ j = 1, \dots, p\Big\} \quad n = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}, \quad f_{ij} = \frac{n_{ij}}{n}$$

Valores numéricos o marcas de clase

Momentos conjuntos no centrados

(centrados en el origen)

$$m_{rs} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} x_i^r\ y_j^s, \quad r, s \in \mathbb{N} \cup \{0\}\ ^{(*)}$$

$\downarrow$

Momentos no centrados marginales:

- $m_{r0} = \sum_{i=1}^{k} f_{i.} x_i^r$
- $m_{0s} = \sum_{j=1}^{p} f_{.j} y_j^s$

Momentos conjuntos centrados

(centrados en medias)

$$\mu_{rs} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(x_i - \overline{x})^r\ (y_j - \overline{y})^s, \quad r, s \in \mathbb{N} \cup \{0\}\ ^{(*)}$$

$\downarrow$

Momentos centrados marginales:

- $\mu_{r0} = \sum_{i=1}^{k} f_{i.}(x_i - \overline{x})^r$
- $\mu_{0s} = \sum_{j=1}^{p} f_{.j}(y_j - \overline{y})^s$

Momentos de primer y segundo orden:

- *Medias marginales*: $m_{10} = \overline{x}, \quad m_{01} = \overline{y}$.
- *Varianzas marginales*: $\mu_{20} = \sigma_x^2, \quad \mu_{02} = \sigma_y^2$.
- *Covarianza de las variables X e Y*: $\mu_{11} = \sigma_{xy} = \mathrm{Cov}(X, Y)$.

Propiedades de los momentos:

- $\sigma_{xy} = \mu_{11} = m_{11} - m_{10} m_{01}$.
- X e Y estadísticamente independientes $\Longrightarrow \begin{cases} m_{rs} = m_{r0} m_{0s} \\ \mu_{rs} = \mu_{r0}\mu_{0s} \\ Cov(X, Y) = \sigma_{xy} = 0. \end{cases}$

$^{(*)}$ Convenio: $0^0 = 1$

PROBLEMA DE REGRESIÓN

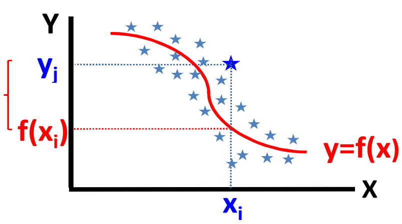
$$\boxed{\begin{array}{c}\left\{((x_i, y_j), n_{ij});\ i = 1, \ldots, k,\ j = 1, \ldots, p\right\} \quad n = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}, \quad f_{ij} = \dfrac{n_{ij}}{n} \\ \text{Valores numéricos o marcas de clase}\end{array}}$$

REGRESIÓN DE Y SOBRE X

Determinar una función, f, que permita aproximar, con el menor error posible, los valores de la variable Y a partir de los valores de X, mediante la variable $f(X)$.

- Y : variable dependiente, explicada, endógena o respuesta
- X : variable independiente, explicativa, exógena o regresora

Residuos $\rightarrow r_{ij}^{y} = y_j - f(x_i)$



CRITERIO DE MÍNIMOS CUADRADOS

Error cuadrático medio asociado a f $\longrightarrow$ $ECM(f) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\ (r_{ij}^{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\ (y_j - f(x_i))^2$

Regresión mínimo cuadrática $\longrightarrow$ Buscar f tal que $ECM(f)$ es mínimo.

CURVAS DE REGRESIÓN MÍNIMO CUADRÁTICAS

La curva de regresión (mínimo cuadrática) de Y sobre X (similarmente, la de X sobre Y) es la que minimiza el error cuadrático medio asociado:

$$Y/X \to f(x_1),\ldots,f(x_k) \;/\; \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - f(x_i))^2 \text{ es mínimo} \Longrightarrow f(x_i) = \overline{y}_i = \sum_{j=1}^{p} f_{j/i}y_j,\ i=1,\ldots,k.$$

$$X/Y \to g(y_1),\ldots,g(y_p) \;/\; \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(x_i - g(y_j))^2 \text{ es mínimo} \Longrightarrow g(y_j) = \overline{x}_j = \sum_{i=1}^{k} f_{i/j}x_i,\ j=1,\ldots,p.$$

$$\Downarrow$$

CURVA DE REGRESIÓN DE Y SOBRE X			
Puntos:	$(x_1,\overline{y}_1)$	$\ldots$	$(x_k,\overline{y}_k)$
	$\downarrow$		$\downarrow$
Frecuencias:	$n_{1.}$		$n_{k.}$

$$\downarrow$$

Residuos mínimo cuadráticos

$r_{ij}^{y} = y_j - \overline{y}_i,\ i=1,\ldots,k;\ j=1,\ldots,p$

CURVA DE REGRESIÓN DE X SOBRE Y			
Puntos:	$(\overline{x}_1,y_1)$	$\ldots$	$(\overline{x}_p,y_p)$
	$\downarrow$		$\downarrow$
Frecuencias:	$n_{.1}$		$n_{.p}$

$$\downarrow$$

Residuos mínimo cuadráticos

$r_{ij}^{x} = x_i - \overline{x}_j,\ i=1,\ldots,k;\ j=1,\ldots,p$

Propiedades de las curvas de regresión mínimo cuadráticas (Y/X)

- *La media de los valores ajustados, $\{(\overline{y}_i, n_{i.});\ i=1,\ldots,k\}$, coincide con la media de los valores observados:*

$$\sum_{i=1}^{k} f_{i.}\overline{y}_i = \overline{y}.$$

- *Los residuos mínimo cuadráticos, $\{(r_{ij}^{y}, n_{ij});\ i=1,\ldots,k,\ j=1,\ldots,p\}$, tienen media cero y, por tanto, su varianza es el error cuadrático medio asociado a la curva de regresión, que coincide con la media de las varianzas condicionadas*:

$$\overline{r}^{y} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}r_{ij}^{y} = 0$$

$$\sigma_{ry}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(r_{ij}^{y})^2 = \sum_{i=1}^{k} f_{i.}\sigma_{y,i}^2 \longrightarrow ECM \text{ asociado a la curva.}$$

- *La varianza de los valores observados es la suma de la varianza de los valores ajustados y la varianza de los residuos*:

$$\sigma_y^2 = \sigma_{ey}^2 + \sigma_{ry}^2, \quad \sigma_{ey}^2 = \sum_{i=1}^{k} f_{i.}(\overline{y}_i - \overline{y})^2.$$

RECTAS DE REGRESIÓN MÍNIMO CUADRÁTICAS

La recta de regresión (mínimo cuadrática) de Y sobre X (y, análogamente, la de X sobre Y) es, entre todas las rectas, la que minimiza el error cuadrático medio:

$$Y/X \rightarrow y = ax + b \ / \ \psi(a,b) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j - (ax_i + b)]^2 \text{ es mínimo}$$

$$\left.\begin{array}{l} \dfrac{\partial \psi(a,b)}{\partial a} = 0 \\[2ex] \dfrac{\partial \psi(a,b)}{\partial b} = 0 \end{array}\right\} \longrightarrow \begin{array}{l} \textit{Ecuaciones normales} \\ m_{11} = am_{20} + bm_{10} \\ m_{01} = am_{10} + b \end{array} \Longrightarrow a = \frac{\sigma_{xy}}{\sigma_x^2}, \quad b = \overline{y} - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x}.$$

$$X/Y \rightarrow x = cy + d \ / \ \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[x_i - (cy_j + d)]^2 \text{ es mínimo} \Longrightarrow c = \frac{\sigma_{xy}}{\sigma_y^2}, \quad d = \overline{x} - \frac{\sigma_{xy}}{\sigma_y^2}\overline{y}.$$

RECTA DE REGRESIÓN DE Y SOBRE X
$y = \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})$
Coeficiente de regresión lineal: $\gamma_{y/x} = \frac{\sigma_{xy}}{\sigma_x^2}$

$\downarrow$

Residuos lineales mínimo cuadráticos
$r_{ij}^{Ly} = y_j - (\overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x_i - \overline{x}))$

RECTA DE REGRESIÓN DE X SOBRE Y
$x = \overline{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \overline{y})$
Coeficiente de regresión lineal: $\gamma_{x/y} = \frac{\sigma_{xy}}{\sigma_y^2}$

$\downarrow$

Residuos lineales mínimo cuadráticos
$r_{ij}^{Lx} = x_i - (\overline{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y_j - \overline{y}))$

Propiedades de las rectas de regresión mínimo cuadráticas (Y/X)

- *Ambas rectas pasan por el centro de gravedad de la distribución,* $(\overline{x}, \overline{y})$.
- *La media de los valores ajustados,* $\{(\overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x_i - \overline{x}), n_{i.}); \ i = 1, \dots, k\}$, *coincide con la media de los valores observados:*

$$\sum_{i=1}^{k} f_{i.}\left(\overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x_i - \overline{x})\right) = \overline{y}.$$

- *Los residuos lineales mínimo cuadráticos,* $\{(r_{ij}^{Ly}, n_{ij}); \ i = 1, \dots, k, \ j = 1, \dots, p\}$, *tienen media cero y, por tanto, su varianza es el error cuadrático medio asociado a la recta:*

$$\overline{r}^{Ly} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} r_{ij}^{Ly} = 0 \Rightarrow \sigma_{ry^L}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(r_{ij}^{Ly})^2 \equiv ECM \textit{ asociado a la recta.}$$

- *La varianza de los valores observados es la suma de la varianza de los valores ajustados,* σ_{xy}^2/σ_x^2, *y la varianza de los residuos:*

$$\sigma_y^2 = \sigma_{ey^L}^2 + \sigma_{ry^L}^2, \qquad \sigma_{ey^L}^2 = \sum_{i=1}^{k} f_{i.}\left(\frac{\sigma_{xy}}{\sigma_x^2}(x_i - \overline{x})\right)^2 = \frac{\sigma_{xy}^2}{\sigma_x^2}.$$

OTROS AJUSTES MÍNIMO CUADRÁTICOS

- **Ajuste de un polinomio de grado arbitrario**: $y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$

$$a_0, \ldots, a_n \;/\; \psi(a_0, \ldots, a_n) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j - (a_0 + a_1x_i + a_2x_i^2 + \ldots + a_nx_i^n)]^2 \text{ es mínimo}$$

$$\left.\begin{array}{l} \dfrac{\partial\psi(a_0,\ldots,a_n)}{\partial a_0} = 0 \\ \dfrac{\partial\psi(a_0,\ldots,a_n)}{\partial a_1} = 0 \\ \qquad\vdots \\ \dfrac{\partial\psi(a_0,\ldots,a_n)}{\partial a_n} = 0 \end{array}\right\} \longrightarrow \begin{array}{rcl} m_{01} & = & a_0 + a_1m_{10} + a_2m_{20} + \ldots + a_nm_{n0} \\ m_{11} & = & a_0m_{10} + a_1m_{20} + a_2m_{30} + \ldots + a_nm_{n+1,0} \\ m_{21} & = & a_0m_{20} + a_1m_{30} + a_2m_{40} + \ldots + a_nm_{n+2,0} \\ \vdots & & \vdots \\ m_{n1} & = & a_0m_{n0} + a_1m_{n+1,0} + a_2m_{n+2,0} + \ldots + a_nm_{n+n,0}. \end{array}$$

- **Ajuste de una hipérbola equilátera**: $y = a/x + b$

  Se invierten los valores de X, $x'_i = 1/x_i$; se ajusta una recta a los nuevos datos, $y = ax' + b$, y se deshace el cambio.

- **Ajuste de una función potencial** $(X, Y > 0)$: $y = bx^a \;\rightarrow\; \log y = a\log x + \log b$

  Se transforman los valores de X e Y tomando logaritmos, $x'_i = \log x_i$, $y'_j = \log y_j$; se ajusta una recta a los nuevos datos, $y' = ax' + b'$, y se deshace el cambio.

- **Ajuste de una función exponencial** $(Y > 0)$: $y = ba^x \;\rightarrow\; \log y = x\log a + \log b$

  Se transforman los valores de Y tomando logaritmos, $y'_j = \log y_j$; se ajusta una recta a los nuevos datos, $y' = a'x + b'$, y se deshace el cambio.

RAZONES DE CORRELACIÓN

Cuantifican el grado en que una variable depende de otra, midiendo el grado de ajuste de la correspondiente curva de regresión a la nube de puntos.

Razón de correlación Y/X

$$\eta^2_{y/x} = \frac{\sigma^2_{ey}}{\sigma^2_y} = 1 - \frac{\sigma^2_{ry}}{\sigma^2_y}$$

- $\sigma^2_y = \sigma^2_{ey} + \sigma^2_{ry} = \sum_{j=1}^{p} f_{.j}(y_j - \overline{y})^2$
- $\sigma^2_{ey} = \sum_{i=1}^{k} f_{i.}(\overline{y}_i - \overline{y})^2$
- $\sigma^2_{ry} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - \overline{y}_i)^2 = ECM(Y/X)$

proporción de σ^2_y debida a la regresión

$$\eta^2_{y/x} = 1 - \frac{ECM(Y/X)}{\sigma^2_y}$$

$$ECM(Y/X) = \sigma^2_y(1 - \eta^2_{y/x})$$

Razón de correlación X/Y

$$\eta^2_{x/y} = \frac{\sigma^2_{ex}}{\sigma^2_x} = 1 - \frac{\sigma^2_{rx}}{\sigma^2_x}$$

- $\sigma^2_x = \sigma^2_{ex} + \sigma^2_{rx} = \sum_{i=1}^{k} f_{i.}(x_i - \overline{x})^2$
- $\sigma^2_{ex} = \sum_{j=1}^{p} f_{.j}(\overline{x}_j - \overline{x})^2$
- $\sigma^2_{rx} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(x_i - \overline{x}_j)^2 = ECM(X/Y)$

proporción de σ^2_x debida a la regresión

$$\eta^2_{x/y} = 1 - \frac{ECM(X/Y)}{\sigma^2_x}$$

$$ECM(X/Y) = \sigma^2_x(1 - \eta^2_{x/y})$$

Propiedades:

- Medidas adimensionales, invariantes frente a cambios de escala y origen en la unidad de medida de la correspondiente variable:

$$\eta^2_{ay+b/x} = \eta^2_{y/x}, \qquad \eta^2_{ax+b/y} = \eta^2_{x/y}.$$

- $0 \leq \eta^2_{y/x},\ \eta^2_{x/y} \leq 1.$

- $\eta^2_{y/x} = 0 \Leftrightarrow \sigma^2_{ey} = 0 \Leftrightarrow \sigma^2_{ry} = \sigma^2_y \Leftrightarrow \overline{y}_1 = \cdots = \overline{y}_k = \overline{y}.$
  $\eta^2_{x/y} = 0 \Leftrightarrow \sigma^2_{ex} = 0 \Leftrightarrow \sigma^2_{rx} = \sigma^2_x \Leftrightarrow \overline{x}_1 = \cdots = \overline{x}_p = \overline{x}.$

- $\eta^2_{y/x} = 1 \Leftrightarrow \sigma^2_{ey} = \sigma^2_y \Leftrightarrow \sigma^2_{ry} = 0 \Leftrightarrow$ Y depende funcionalmente de X.
  $\eta^2_{x/y} = 1 \Leftrightarrow \sigma^2_{ex} = \sigma^2_x \Leftrightarrow \sigma^2_{rx} = 0 \Leftrightarrow$ X depende funcionalmente de Y.

MEDIDAS DE CORRELACIÓN LINEAL

Cuantifican el grado de relación lineal entre las variables, midiendo el grado de ajuste de las rectas de regresión a la nube de puntos.

Coeficiente de determinación lineal

$$R_{xy}^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\ \sigma_y^2}$$

proporción de σ_x^2 y de σ_y^2 debida a la regresión lineal

Coeficiente de correlación lineal

$$R_{xy} = \frac{\sigma_{xy}}{\sigma_x\ \sigma_y}$$

$$R_{xy}^2 = 1 - \frac{ECM^L(Y/X)}{\sigma_y^2} = 1 - \frac{ECM^L(X/Y)}{\sigma_x^2}$$

$$ECM^L(Y/X) = \sigma_y^2(1 - R_{x,y}^2); \quad ECM^L(X/Y) = \sigma_x^2(1 - R_{x,y}^2).$$

Propiedades:

- $R_{xy}^2 = \gamma_{y/x}\gamma_{x/y}$.
- Medidas adimensionales, invariantes frente a cambios de escala (salvo el signo de R_{xy}) y origen en la unidad de medida de cualquiera de las variables:
$$R_{ax+b,cy+d}^2 = R_{x,y}^2.$$
- $R_{xy}^2 \leq \eta_{x/y}^2, \ \ R_{xy}^2 \leq \eta_{y/x}^2$.
- $0 \leq R_{xy}^2 \leq 1, \quad -1 \leq R_{xy} \leq 1$.
- $R_{xy}^2 = 0\ (R_{xy} = 0) \Leftrightarrow \sigma_{xy} = 0 \ \Leftrightarrow$ las rectas de regresión son $y = \overline{y},\ x = \overline{x}$.
- $R_{xy}^2 = 1\ (R_{xy} = \pm 1) \Leftrightarrow$ existe dependencia lineal (recíproca) entre X e Y.
- $0 < R_{xy} \leq 1 \ \longrightarrow$ rectas de regresión crecientes.
- $-1 \leq R_{xy} < 0 \ \longrightarrow$ rectas de regresión decrecientes.