Tema 6

Variables estadísticas bidimensionales. Regresión lineal

Bidimensionales. Regresión lineal

- Variables bidimensionales:
  - Frecuencias
  - Tablas
  - Gráfico
- Grado de dependencia. Independencia
- Covarianza y correlación
- Idea de la regresión
- Regresión lineal:
  - Lenguaje de regresión
  - Criterios de ajuste
  - Rectas de regresión
  - Bondad de ajuste

Variables bidimensionales

- Muchas veces se busca estudiar la **relación** que existe entre **dos características** estudiadas sobre la **misma población: variable bidimensional (X,Y).**
- **N** es el número de veces que se repite la experiencia en la que se observan las dos características.
- Los resultados diferentes obtenidos en las N experiencias son los pares (x_i, y_j)

$$i = 1, 2, ..., k \quad j = 1, 2, ..., m.$$

Frecuencias

- La ***frecuencia absoluta*** de cada par (x_i, y_j) se representará por $\mathbf{n_{ij}}$ o $\mathbf{n(x_i,y_j)}$, y es el **número** de veces que aparece dicho par.
- La ***frecuencia relativa*** de cada pareja (x_i, y_j) se representará por $\mathbf{f_{ij}}$ o $\mathbf{f(x_i,y_j)}$, y es la **proporción** de veces que aparece dicho par en las N experiencias: $$f(x_i, y_j) = \frac{n(x_i, y_j)}{N} \quad \forall i, j.$$
- Se cumple: $$\sum_{i=1}^{k}\sum_{j=1}^{m} n(x_i, y_j) = N \quad \text{y} \quad \sum_{i=1}^{k}\sum_{j=1}^{m} f(x_i, y_j) = 1.$$

Tabla estadística una entrada

- A veces se manejan tablas de **una entrada** (como en **la entrada de datos a una calculadora**):

x_i	x_1	x_1	⋯	x_1	x_2	x_2	⋯	x_2	⋯	x_k	x_k	⋯	x_k
y_j	y_1	y_2	⋯	y_m	y_1	y_2	⋯	y_m	⋯	y_1	y_2	⋯	y_m
n_{ij}	n_{11}	n_{12}	⋯	n_{1m}	n_{21}	n_{22}	⋯	n_{2m}	⋯	n_{k1}	n_{k2}	⋯	n_{km}
f_{ij}	f_{11}	f_{12}	⋯	f_{1m}	f_{21}	f_{22}	⋯	f_{2m}	⋯	f_{k1}	f_{k2}	⋯	f_{km}

Tabla estadística de dos entradas

- En otras ocasiones (como en tablas de contingencia) se construyen **tablas estadísticas de doble entrada.** La siguiente es una tabla con frecuencias absolutas:

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}
$\vdots$					$\ddots$	
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{km}

Diagrama de dispersión

- El **diagrama de dispersión o nube de puntos** es la **representación gráfica** más habitual para las variables bidimensionales.
- Para ello, en unos ejes coordenados, se marcan los puntos de coordenadas (x_i, y_j), indicando cuál es su frecuencia absoluta (no es necesario indicarla cuando su frecuencia es 1).

Grado de dependencia entre X e Y

En la variable bidimensional (X, Y) las variables X e Y tienen una relación más o menos fuerte.

Esta relación varía entre dos casos extremos:

- **Independencia:** el comportamiento de la variable X no influye en el de la variable Y, y viceversa.
- **Dependencia funcional:** la relación entre ambas variables es tan estrecha que existe una función f, de forma que Y=f(X) (o bien existe g, de forma que X=g(Y)).

En la mayor parte de las situaciones nos encontraremos en un grado de dependencia intermedio entre la independencia y la relación funcional.

Variables independientes

- Las variables X e Y son independientes si:

$$f(x_i, y_j) = f(x_i) \cdot f(y_j) \quad \forall\, i, j.$$

- Ejemplo: en la distribución de frecuencias de la tabla siguiente, X e Y son **independientes**:

X\Y	-2	-1	0	3	n_i
3	2	5	3	1	11
4	6	15	9	3	33
5	4	10	6	2	22
n_j	12	30	18	6	66

Nota: en variables independientes las **filas** de frecuencias son **proporcionales** entre sí (lo mismo le pasa a las **columnas**).

Covarianza

- La **covarianza** de las variables X e Y, S_{XY}, es una **medida de la relación entre X e Y**. Se calcula como:

$$S_{XY} = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{m} (x_i - \overline{X})(y_i - \overline{Y}) \cdot n(x_i, y_j).$$

  - Si X e Y son independientes, entonces $S_{XY} = 0$ (**el recíproco no es cierto**).
  - **Si $S_{XY}>0$,** entonces entre X e Y hay dependencia directa (a **mayor** valor de X, **mayor** valor de Y, o viceversa).
  - **Si $S_{XY}<0$,** entonces entre X e Y hay dependencia inversa (a **mayor** valor de X, **menor** valor de Y, o viceversa).

Coeficiente de correlación

- Se define el coeficiente de **correlación lineal de Pearson** entre dos variables X e Y, y se denota por **r**, a

$$r = \frac{S_{XY}}{S_X S_Y}.$$

  - $-1 \leq r \leq 1$.
  - **Si X e Y son independientes, entonces r = 0 (el recíproco no es cierto).**
  - **Si r > 0** entre X e Y hay dependencia directa (a **mayor** valor de X, **mayor** valor de Y, o viceversa).
  - **Si r <0** entre X e Y hay dependencia inversa (a **mayor** valor de X, **menor** valor de Y, o viceversa).

Idea de la regresión

- En la regresión se plantea expresar la variable Y como función de X: Y=f(X) (o X como función de Y: X=g(Y)).
- Esto, en general, no es posible.
- En la mayoría de los casos lo que se va a conseguir es calcular **una función que exprese Y en función de X con el menor error posible**: regresión de Y sobre X (o X en función de Y: regresión de X sobre Y).

Regresión más simple

- Nos conformaremos con ajustar a una **recta** (que es la función **más sencilla**): regresión lineal.

- También se podrían realizar ajustes a otro tipo de funciones más complejas: regresión no lineal.

  No entraremos en este tipo de regresión.

Lenguaje de regresión

- En la **regresión de Y sobre X**:
  - X es la **variable independiente o explicativa**.
  - Y variable **dependiente o a predecir**.
- Si se trata de predecir un valor de la variable dependiente cuando la variable explicativa toma un valor en el intervalo $[x_1,x_k]$, se dice que la predicción es una **interpolación**. En caso contrario es una **extrapolación.**
- En general, las **interpolaciones son fiables** cuando se tiene un "buen ajuste", mientras que las extrapolaciones no suelen ser fiables.

Regresión lineal

- En la regresión lineal se va a encontrar la recta

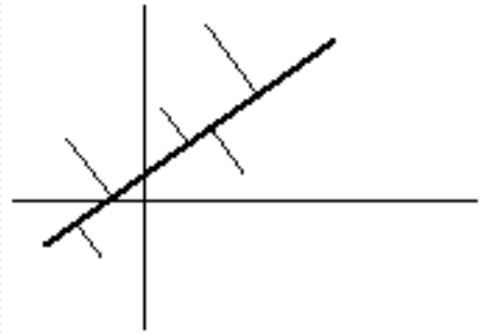
$$y = a + bx$$

que mejor se "ajusta" a todos los puntos (x_i, y_j) de la variable bidimensional (X, Y): **recta de regresión de Y sobre X**.

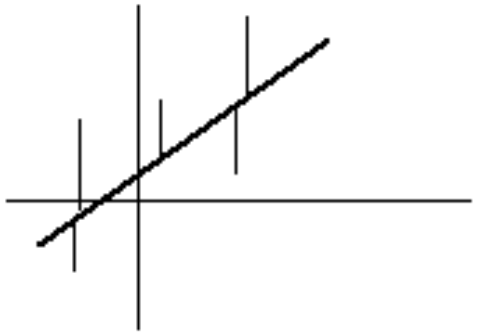
- Si se intercambian los papeles de X e Y se obtiene la recta de regresión de X sobre Y.

Criterios de ajuste

- **¿Qué criterio debe ser utilizado para** "acoplar" o "ajustar" **la recta a la nube de puntos?**



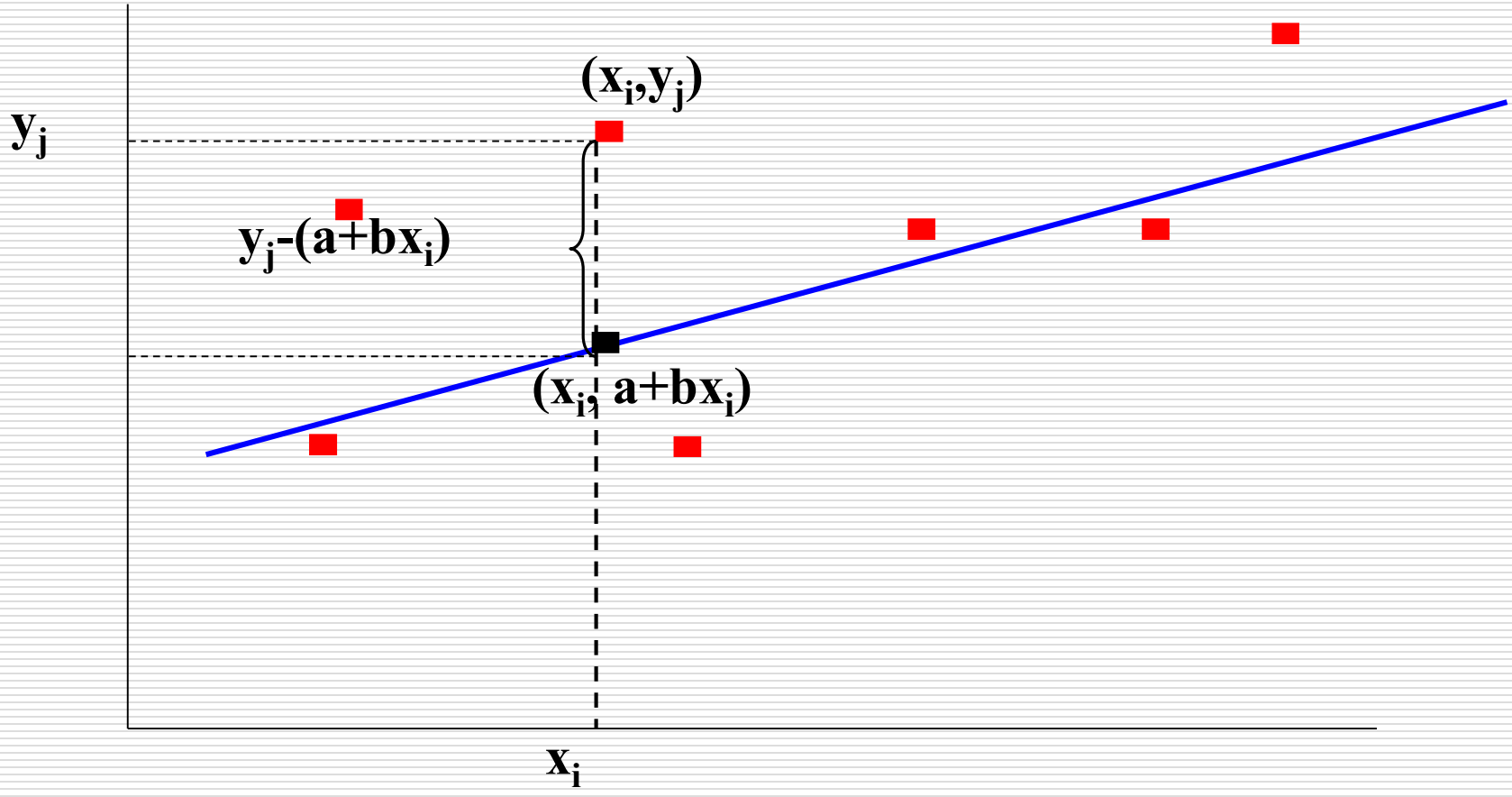
Modelo de regresión ortogonal



Modelo de regresión mínimo-cuadrática

- Por **operatividad** se elige el modelo de regresión mínimo-cuadrático.

Regresión mínimo-cuadrática



Regresión lineal mínimo-cuadrática

- Las **separaciones elevadas al cuadrado** son **ponderadas** por la frecuencia asociada, con lo que **el objetivo** es ***minimizar***:

$$\overline{E^2}(a,b) = \sum_{i=1}^{k}\sum_{j=1}^{m} f_{ij}\left(y_j - a - bx_i\right)^2$$

- Resolviendo se obtienen los valores:

$$a = \overline{Y} - \frac{S_{XY}}{S_X^2}\overline{X}, \qquad b = \frac{S_{XY}}{S_X^2}.$$

Recta de regresión de Y/X

- Sustituyendo los valores de a y de b en la recta **y = a + b x**, obtenemos la expresión:

$$y - \overline{Y} = \frac{S_{XY}}{S_X^2}\left(x - \overline{X}\right)$$

que es la ***recta de regresión de Y sobre X.***

- Intercambiando los papeles de X e Y se obtiene la ***recta de regresión de X sobre Y*:**

$$x - \overline{X} = \frac{S_{XY}}{S_Y^2}\left(y - \overline{Y}\right)$$

Rectas de regresión

- La recta de regresión de Y/X y la de X/Y, en general no coinciden, se cortan en el punto:

$$\left(\overline{X}, \overline{Y}\right)$$

y ambas tienen **pendientes del mismo signo** (el de la covarianza o el del coeficiente de correlación).

Interpretación pendiente r.r. Y/X

- En la recta de **regresión de Y sobre X** la pendiente es el coeficiente "b",

$$b = \frac{S_{XY}}{S_X^2},$$

y se interpreta como **la variación que experimenta la variable Y cuando X varía en una unidad.**

Interpretación pendiente r.r. X/Y

- En la recta de **regresión de X sobre Y** la pendiente es el **inverso** del coeficiente "b´",

$$b' = \frac{S_{XY}}{S_Y^2},$$

y se interpreta como **la variación que experimenta la variable X cuando Y varía en una unidad** .

Variable predicción

- La variable

$$Y^* = a + bX = \overline{Y} + \frac{S_{XY}}{S_X^2}\left(X - \overline{X}\right)$$

es la variable predicción de Y a partir de X en la recta de regresión de Y sobre X.

- Sus valores son las predicciones

$$y_i^* = f\left(x_i\right) = a + bx_i = \overline{Y} + \frac{S_{XY}}{S_X^2}\left(x_i - \overline{X}\right)$$

- Intercambiando X e Y se tiene la predicción de X a partir de Y.

Variable error

- La variable

$$E = Y - Y^* = Y - \left(\overline{Y} + \frac{S_{XY}}{S_X^2}\left(X - \overline{X}\right) \right)$$

  es la variable error al predecir Y a partir de X con la recta de regresión de Y sobre X.

- Sus valores son los errores

$$e_{ij} = y_j - y_i^* = y_j - \left(\overline{Y} + \frac{S_{XY}}{S_X^2}\left(x_i - \overline{X}\right) \right).$$

- Intercambiando X e Y se tienen los errores de X a partir de Y.

Bondad del ajuste lineal

- Tras encontrar la recta que mejor se "acopla" a un conjunto de datos bidimensionales, hay que medir **el "grado de acoplamiento"** (bondad del ajuste) de los datos a la función.
- El caso ideal es cuando todos los puntos están sobre la recta (dependencia funcional). En general esto **no ocurre**.
- Por tanto, hay que "medir" el grado de alejamiento de los datos de la muestra a la recta de regresión calculada (**o variabilidad de los datos respecto a la recta**).

Coeficiente r y r^2

- Los coeficientes de bondad del ajuste lineal más utilizados son:
  - r, ***coeficiente de correlación*** entre las variables X e Y :
    $$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$
  - El cuadrado de este coeficiente, $\mathbf{r^2}$, recibe el nombre de ***coeficiente de determinación.*** También se puede calcular como
    $$r_{XY}^2 = \frac{S_{y^*}^2}{S_Y^2} = 1 - \frac{S_E^2}{S_Y^2}.$$

Interpretación de r^2

$r^2 \times 100$ es el porcentaje de variabilidad de Y que queda explicada por la regresión.

- Si **$r^2 = 1$:** el 100% la variabilidad de Y queda explicada por la regresión. El **ajuste es perfecto** (los puntos están sobre la recta de regresión). Ambas rectas de regresión **coinciden.**

- Si **$r^2 = 0$** (variables ***incorreladas***) nada en la variabilidad de Y está explicada por la regresión lineal a partir de X **(independientes linealmente**). Las rectas de regresión son **perpendiculares** y paralelas a los ejes.

- **Suele considerarse que la regresión es aceptable si $r^2 \geq 0.75$.**

Interpretación de $1-r^2$

$(1-r^2)\times 100$ es el porcentaje de variabilidad de Y que no está explicada por la regresión.

- Si $1-r^2 = 0$: el 0% la variabilidad de Y no está explicada por la regresión. El **ajuste es perfecto** (los puntos están sobre la recta de regresión). Ambas rectas de regresión **coinciden.**
- Si $1-r^2 = 1$ (variables ***incorreladas***) el 100% de variabilidad de Y está sin explicar por la regresión lineal a partir de X **(independientes linealmente**). Las rectas de regresión son **perpendiculares** y paralelas a los ejes.
- **Suele considerarse que la regresión es aceptable si el porcentaje no explicado es $1-r^2<0.25$.**

Fiabilidad de la predicción

- Recuerda que la predicción puede ser **interpolación** o **extrapolación**.
  - Las extrapolaciones no suelen ser fiables.
  - Las interpolaciones son fiables si **$r^2 \geq 0.75$** (ya que el ajuste es aceptable en este caso).