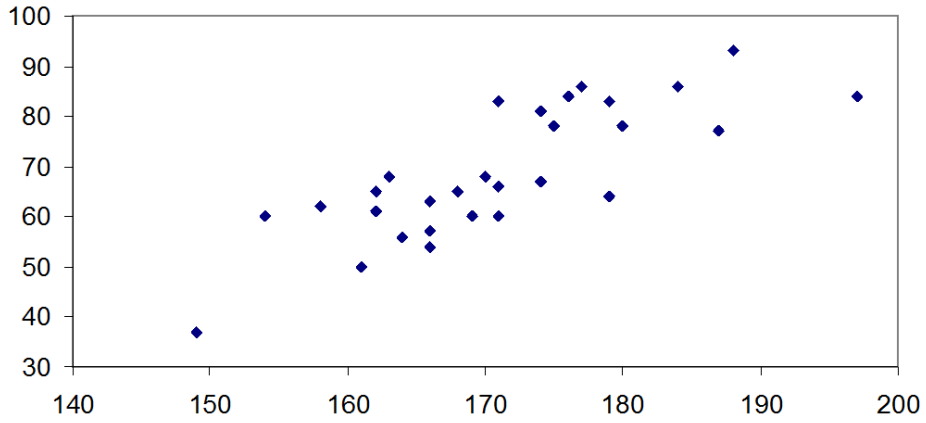
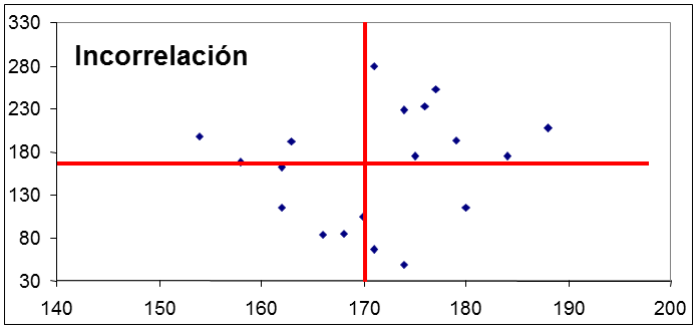
1 DEFINICIONES PREVIAS

- **Regresión**: implica la obtención de una ecuación mediante la que podamos estimar el valor medio de una variable.
- **Correlación**: es la cuantificación del grado de relación existente entre dos variables.
- **Nube de puntos o diagrama de dispersión**: es un gráfico donde se representan los pares ordenados (x,y). Nos ayuda a determinar si es la regresión lineal es aplicable. Si lo es, los puntos deberán mostrar una notable tendencia a la linealidad.

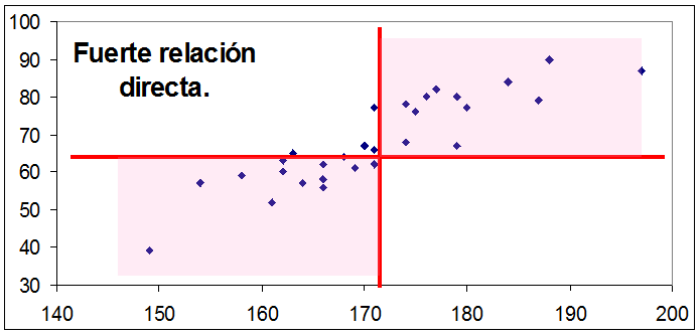


- Nuestro problema será estimar matemáticamente la ecuación de la recta que mejor ajusta los datos

2 RELACIÓN DIRECTA E INVERSA



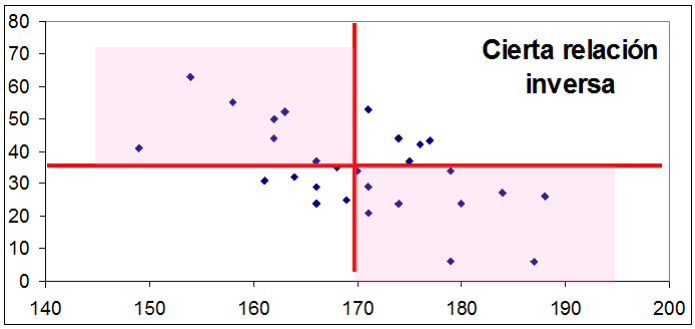
Incolerrelación: para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares.



Relación directa:

Para los valores de X mayores que la media le corresponden valores de Y mayores también.

Para los valores de X menores que la media le corresponden valores de Y menores también.



Relación inversa:

Para los valores de X mayores que la media le corresponden valores de Y menores.

3 COVARIANZA

Es una medida que nos informará de la variabilidad conjunta de dos variables cuantitativas. Se calcula con cualquiera de las dos fórmulas siguientes:

$$\sigma_{xy} = \frac{1}{N} \cdot \sum_{i=0}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$\sigma_{xy} = \frac{1}{N} \cdot \sum_{i=0}^{N} x_i \cdot y_i - \bar{x} \cdot \bar{y}$$

- Si la covarianza es positiva ambas variables crecerán o decrecerán a la vez (relación directa).
- Si la covarianza es negativa cuando una de las variables crece la otra decrece (relación inversa).
- Si la covarianza es cero los puntos se distribuyen uniformemente en los cuatro cuadrantes y hay poca correlación lineal entre las variables (incorreladas). Ojo puede existir relación entre las variables, una covarianza nula no implica independencia de las variables.
- El signo de la covarianza nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el grado de relación entre las variables.

4 COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON, r

Es el parámetro que nos informa sobre la fuerza de la correlación o dependencia lineal en una lista de valores (x_1,y_1), (x_2,y_2),...,(x_n,y_n) de una variable bidimensional. Se calcula mediante la fórmula:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

A σ_x y σ_y se las llama desviaciones típicas marginales y se calculan:

$$\sigma_x = \sqrt{\frac{1}{N} \cdot \sum_{i=0}^{N} (x_i - \bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{N} \cdot \sum_{i=0}^{n} (y_i - \bar{y})^2}$$

4.1 PROPIEDADES DE r

- Es adimensional.
- Siempre toma valores entre -1 y 1.
- Si las variables están tipificadas r = σ_{xy}.
- Cuando el valor de r se acerca a 1 o a -1 la correlación lineal es fuerte. Por el contrario, valores de r próximos a cero indican una escasa correlación lineal.
- Si r tiene signo positivo significa que existe correlación positiva, esto es, la recta en torno a la que se agrupan los puntos tiene pendiente positiva. Si el signo de r es negativo, entonces la recta tiene pendiente negativa.

5 RECTA DE REGRESIÓN

Si X e Y son dos variables cuantitativas, la recta de regresión de Y sobre X se obtiene:

$$y - \bar{y} = r \cdot \frac{\sigma_y}{\sigma_x} \cdot (x - \bar{x})$$

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} \cdot (x - \bar{x})$$

Esta recta se puede representar de la forma:

$$y = b \cdot x + a \qquad b = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Esta recta se llama también recta de regresión mínimo cuadrática debido a que es la que hace mínima la suma de los cuadrados de las distancias de los puntos (x,y) a la recta. Es la recta que mejor ajusta los datos.

A la hora de utilizar esta recta para realizar predicciones hemos de tener en cuenta que:

- La predicción efectuada (y) y el valor real obtenido (y) rara vez coincidirán. A la cantidad e=y-y se le denomina residuo o error residual.
- Si el valor de r es cero o está próximo a cero, la recta no es muy útil y la predicción será, en general, mala.
- Si el valor de r está muy próximo a uno la estimación será bastante buena siempre y cuando los valores sobre los que hacemos dicha estimación no se encuentren muy alejados del rango de los datos que hemos utilizado para calcular la recta.

  En concreto si los valores sobre los que efectuamos la predicción están dentro del rango de valores de las variables estudiadas la predicción será fiable. Si dichos valores caen fuera del rango de las variables pero cerca, por arriba o por abajo, la predicción será útil pero debemos tomarla con cierta precaución. Por último si los valores sobre los que efectuamos la predicción están fuera del rango de las variables y están muy alejados la predicción no será, en absoluto, fiable.

6 EXTENSIONES DEL MODELO LINEAL

A veces los datos se ajustarán mejor a una curva que a una recta. Existen muchas posibles curvas para ajustar nuestros datos. Nosotros veremos dos posibilidades, además de la recta.

6.1 MODELO EXPONENCIAL

Para explorar la dependencia exponencial de una variable bidimensional, (x,y), se hace una regresión mínimo cuadrática para los datos $(x_1, \ln y_1), (x_2, \ln y_2), \dots, (x_n, \ln y_n)$ y se obtiene la recta $ln\, y = b \cdot x + a$ y despejando nos queda la expresión:

$$y = k \cdot e^{b \cdot x} \qquad \text{donde } k = e^a$$

$$b = \frac{\sigma_{x \ln y}}{\sigma_x^2}$$

$$a = \overline{\ln y} - b \cdot \bar{x}$$

6.2 MODELO POTENCIAL

Para explorar la dependencia potencial de una variable bidimensional, (x,y), se hace una regresión mínimo cuadrática para los datos $(\ln x_1, \ln y_1), (\ln x_2, \ln y_2), \dots, (\ln x_n, \ln y_n)$ y se obtiene la recta $ln\, y = b \cdot ln\, x + a$ y despejando nos queda la expresión:

$$y = k \cdot x^b \qquad \text{donde } k = e^a$$

$$b = \frac{\sigma_{\ln x \ln y}}{\sigma_{\ln x}^2}$$

$$a = \overline{\ln y} - b \cdot \overline{\ln x}$$

EJERCICIO 1. Representa el diagrama de dispersión y calcula las medias, desviaciones típicas, la covarianza y el coeficiente de correlación de los datos: (2,4), (3,6), (6,10).

EJERCICIO 2. En algunos lugares hay una fuerte asociación entre concentraciones de dos contaminantes diferentes. El artículo "The Carbon Component of the Los Angeles Aerosol: Source Apportionment and Contributions to the Visibility Budget" reporta los siguientes datos sobre concentración de ozono, X (en ppm) y concentración de carbono secundario, Y (mg/m^3). Hallar la recta de regresión lineal.

X	0,066	0,088	0,12	0,05	0,162	0,186	0,057	0,1
Y	4,6	11,6	9,5	6,3	13,8	15,4	2,5	11,8

X	0,112	0,055	0,154	0,074	0,111	0,14	0,071	0,11
Y	8	7	20,6	16,6	9,2	17,9	2,8	13

EJERCICIO 3. Se hicieron las mediciones de concentración de hidrógeno (ppm) de las mismas muestras mediante una técnica de cromatografía de gases (X) y mediante un nuevo sensor (Y).

X	Y
47	38
62	62
65	53
70	67
70	84
78	79
95	93
100	106

X	Y
114	117
118	116
124	127
127	114
140	134
140	142
150	170
164	154

- Dibuja el diagrama de dispersión
- Obtén la ecuación de regresión lineal que mejor se ajusta y su coeficiente de correlación.

EJERCICIO 4. En un proceso químico, el tiempo de reacción Y (horas) está relacionado con la temperatura (ºF) de la cámara donde ocurre la reacción según el modelo de regresión lineal con ecuación $Y = 5{,}00 - 0{,}01 \cdot x$. ¿Cuál es el cambio esperado del tiempo de reacción para un aumento de 1 ºF en la temperatura? ¿Y para un aumento de 10 º F?

EJERCICIO 5. Varios estudios han demostrado que los líquenes son excelentes indicadores biológicos de la contaminación del aire. Se tiene la siguiente información sobre el depósito de NO_3 (X) húmedo (g. N /m2) y de liquen (Y) N (% de peso en seco).

X	Y
0,05	0,48
0,10	0,55
0,11	0,48
0,12	0,50
0,31	0,58
0,37	0,52
0,42	1,02

X	Y
0,58	0,86
0,68	0,86
0,68	1,00
0,73	0,88
0,85	1,04
0,92	1,70

- Elije el modelo que mejor ajuste los datos entre el lineal, potencial o exponencial.

EJERCICIO 6. Se ha realizado un ensayo clínico para estudiar el posible efecto hipotensor de un fármaco. Se ha medido la tensión arterial diastólica (TAD) en condiciones basales antes de comenzar el tratamiento y un mes después. Los resultados fueron los siguientes:

TAD previa	96	101	115	104	100	98	94	105	95	100	99	110	99
TAD posterior	84	92	91	89	89	83	87	93	77	91	88	103	82

- ¿Existe relación lineal entre ambas tensiones?
- ¿Cuál es la TAD esperado, tras el tratamiento, en paciente que presentó una TAD basal de 98 mmHg?

EJERCICIO 7. Se han llevado a cabo 12 tomas de presión intracraneal (en mmHg) en animales de laboratorio mediante dos métodos diferentes, el clásico y uno experimental. Los resultados obtenidos por ambos métodos se muestran a continuación. Hallar la recta de regresión.

Clásico	10	12	25	72	32	35	81	74	29	51	56	61
Experimental	8	11	23	66	29	34	78	70	25	48	53	59

- ¿Qué medición de presión intracraneal cabría esperar por el método experimental, si el método clásico ha dado un resultado de 48 mmHg? ¿es fiable la predicción realizada?
- ¿Qué medición de presión intracraneal cabría esperar por el método clásico, si el método experimental ha dado un resultado de 100 mmHg? ¿es fiable la predicción realizada?

EJERCICIO 8. Se han tomado los pesos (en kilos) y longitudes (en metros) de una muestra de serpientes pitón criadas en cautividad. Los resultados se muestran en la tabla siguiente.

Peso	4,5	3,2	5,3	4,7	4,9	5,1	5	5,5	3,9	4,2	4,4	5,5
longitud	4,1	3,7	4,9	4,5	4,3	4,4	4,2	5,1	4,2	4,1	3,9	4,8

- Determinar la curva que mejor ajusta los datos.
- ¿Qué peso podríamos esperar para una serpiente de cuatro metros de longitud? Discute la validez de la previsión realizada.
- ¿Qué longitud diríamos que podría alcanzar una serpiente que pesó 87 kilos? Discute la validez del pronóstico.