Variables estadísticas bidimensionales

En ocasiones los datos se presentan en pares (x, y). Es decir que al recoger los datos no solo tenemos una variable, sino dos y hay una correspondencia entre los valores de una y los de la otra.

Pensemos en el siguente ejemplo:

0.0.1 Ejemplo

Una empresa observa que parece haber una relación fuerte entre las ventas en enero y las ventas en febrero. Para ello se recogen los datos de 9 años y se anotan en una tabla las ventas de enero y febrero de cada año en miles de euros.

ventas enero	ventas febrero
142.74	69.06
146.58	70.62
149.01	72.03
151.72	73.48
154.12	74.89
158.23	76.48
160.19	77.85
165.46	79.54
168.82	81.05

Para analizar estos datos, podríamos trabajar con cada variable por separado (calculando medias, varianzas, haciendo gráficas...), pero se perdería la relación entre ambas. Siguendo este procedimiento sería dificil observar por ejemplo como a mayores ventas en enero corresponden mayores ventas en febrero.

Para ello trabajamos con las dos variables juntas, a través de la **Covarianza** y el **coeficiente de correlación de pearson**. Mediante ellos intentaremos capturar la relación de dependencia entre ambas variables, particularmente la *dependencia lineal,*, que ocurre cuando una de las variables puede ser aproximadas a partir de la otra mediante una recta.

1 Covarianza

La **covarianza** es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal.

Supongamos que tenemos unos datos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ... (x_N, y_N)$$

La covarianza se denota por S_{xy} y se define como

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

$$= \frac{1}{N}((x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + ... + (x_N - \overline{x})(y_N - \overline{y}))$$

donde $\overline{x}$ denota la media de la primera variable (x), e $\overline{y}$ denota la media de la segunda (y)

1.1 Ejemplo

Para entender cómo calcularlo, usaremos el ejemplo anterior. Primero calculamos la media de ambas variables. En este caso como todos los datos son distintos no hay frecuencias así que para calcular las medias basta sumar y dividir entre el número de datos.

Obtenemos $\overline{x} \cong 155.207$, $\overline{y} \cong 77.337$. Añadimos una columna calculando las multiplicaciones $(x_i - \overline{x})(y_i - \overline{y})$

x_i	y_i	$(x_i - \overline{x})(y_i - \overline{y})$
142.74	69.06	80.389
146.58	70.62	21.981
149.01	72.03	16.658
151.72	73.48	4.142
154.12	74.89	-1.916
158.23	76.48	-1.564
160.19	77.85	2.502
165.46	79.54	27.908
168.82	81.05	114.372

y para calcular la covarianza basta sumar esta tercera columna y dividir entre el número de datos

$$S_{xy} \cong 264.474/9 \cong 29.386$$

1.2 El gráfico nubes de puntos

Una manera de visualizar la relación o dependencia entre las dos variables es dibujar cada punto (x_i, y_i) en el plano.

En el ejemplo anterior, el gráfico sería el siguente.

2 Coeficiente de correlación de Pearson

El coeficiente de **correlación de Pearson** es una medida de dependencia lineal entre dos variables estadísticas cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

Se define como

$$\rho_{XY} = \frac{S_{xy}}{S_X S_Y}$$

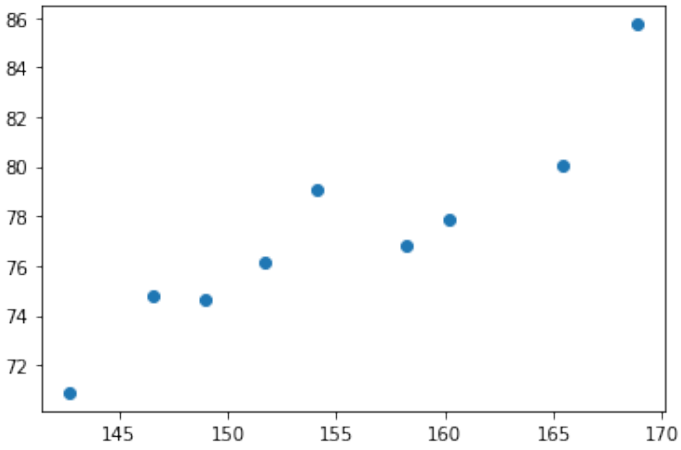


Figure 1: nube de puntos

donde S_{xy} de nota la covarianza, S_X denota la desviación típica de la primera variable y S_Y la desviación típica de la segunda.

en el ejemplo anterior, si calculamos además la desviación típica de X e Y obtenemos

$$S_x \cong 8.205$$
$$S_y \cong 3.920$$

luego

$$\rho_{XY} \cong \frac{29.386}{8.205 \cdot 3.920} \cong 0.913$$

Deducimos de aquí que dado que el coeficiente de correlación de Pearson es cercano a 1 existe una dependencia lineal directa entre las ventas de enero y febrero

2.1 Interpretación del coeficiente de correlación de Pearson

- Si $\rho_{XY} > 0$ hay dependencia lineal directa (positiva), es decir, a grandes valores de X corresponden grandes valores de Y.

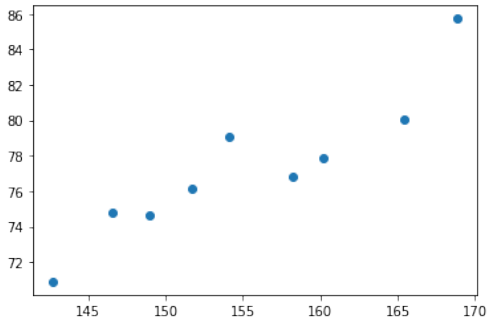


Figure 2: nube de puntos correlación positiva

- Si $\rho_{XY} = 0$ se interpreta como la no existencia de una relación lineal entre las dos variables.

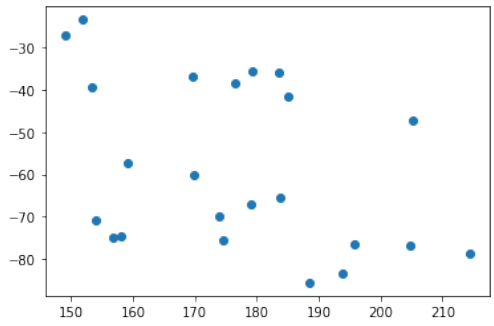


Figure 3: nube de puntos correlación 0

- Si $\rho_{XY} < 0$ hay dependencia lineal inversa o negativa, es decir, a grandes valores de X corresponden pequeños valores de Y.

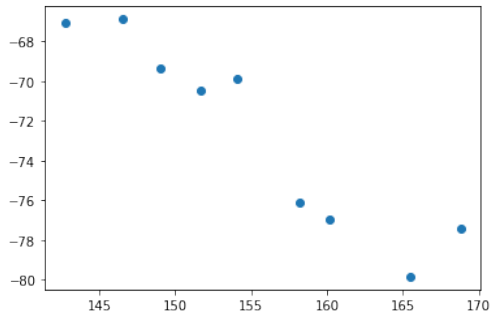


Figure 4: nube de puntos correlación negativa

3 Regresión simple

En las nubes de puntos que hemos utilizado hasta ahora, veíamos como en los casos en que ρ_{XY} es cercano a -1 o 1, los puntos (X, Y) parecen acercarse a una recta que puede ajustarse visualmente. Esta recta es la **recta de regresión**.

En esta sección, aprenderemos a calcular la línea de regresión de manera más precisa, usando la ecuación más sencilla que relaciona las dos variables matemáticamente. Aquí, examinaremos sólo relaciones lineales entre dos variables. Recordemos que la ecuación de una recta viene dada por

$$Y = a + bX$$

Habitualmente, dados unas variables X e Y, será la variable Y la que querremos predecir a partir de la X, por eso llamaremos a Y *variable dependiente* ya a la X *variable independiente*.

3.1 Método de mínimos cuadrados

Imaginemos que tenemos una recta

$$f(X) = a + bX$$

El valor $\hat{Y} = f(X)$ representa el valor con el que intentamos predecir Y. Por lo tanto el error (o residuo) de la predicción es precisamente

$$Y - \hat{Y} = Y - f(X) = Y - a - bX$$

Una manera de trabajar con el error es trabajar con el cuadrado de la expresión anterior, es decir, el cuadrado del error de la predicción

$$(Y - f(X))^2 = (Y - a - bX)^2$$

Si tenemos una muestra

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ... (x_N, y_N)$$

Y un modelo como el anterior $f(X) = a + bY$, los errores medidos de la forma anterior son las distancias en vertical entre la recta y cada punto como puede verse en la figura siguiente:

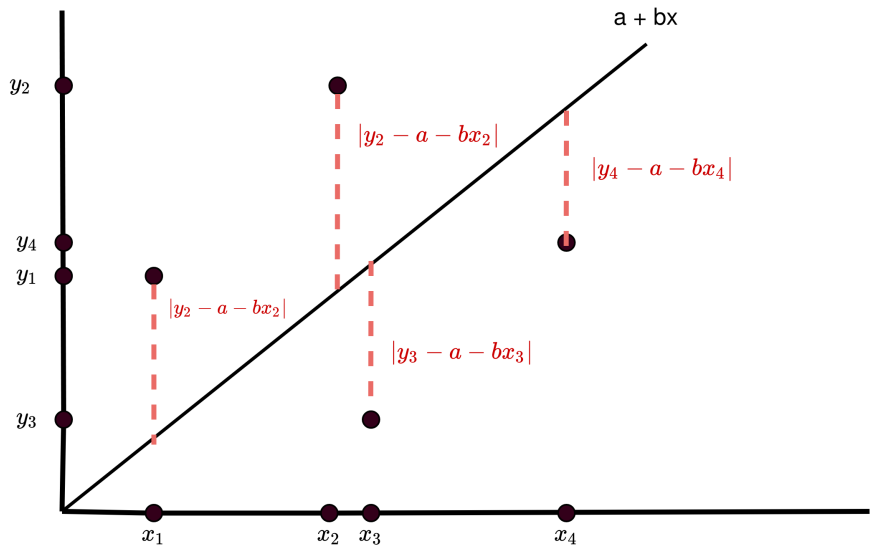


Figure 5: los cuadrados de las distancias son precisamente los errores $(Y - f(X))^2$

De tal modo que podemos decir que el error **global** de aproximar Y con la recta $f(X) = a + bX$ se puede medir como la suma de todos los cuadrados de las distancias anteriores. Pensemos que cuanto más alejados estén los puntos de la recta, *peor* aproxima la recta.

Por lo tanto definimos el **residuo suma de cuadrados** como

$$S(a, b) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2$$

Lo llamamos de esta manera $S(a, b)$ puesto que es el error que se produce al elegir a y b como parámetros de la recta de regresión.

De este modo, los valores de a y b que **menor error de aproximación** $S(a, b)$ produzcan serán los más deseados. Esos valores precisamente son los que definen la **recta de regresión**.

Para encontrar esos valores debemos calcular el mínimo de la función $S(a,b)$

deberemos calcular

$$\frac{\partial S}{\partial a} = 0$$
$$\frac{\partial S}{\partial b} = 0$$

resolver el sistema y averiguar que sea mínimo. Si hacemos esto encontraremos que

$$a = \overline{y} - \frac{S_{XY}}{S_X^2}\overline{x}$$
$$b = \frac{S_{XY}}{S_X^2}$$

Para calcular la estimación $\widehat{y}$ que nuestro modelo produce para un valor de x calculamos

$$\widehat{y} = a + bx$$

3.1.1 Ejemplo

Para el ejemplo con el que comenzamos este tema, en el que teníamos una tabla con las ventas de enero (X) y las de febrero (Y), habiamos calculado

$$S_{xy} \cong 29.386$$
$$S_x \cong 8.205$$
$$S_y \cong 3.920$$
$$\overline{x} \cong 155.207$$
$$\overline{y} \cong 77.337$$

De modo que si queremos calcular la recta de regresión de Y sobre X (es decir usar las ventas en enero para predecir las de febrero, esto es X para predecir Y) tenemos que calcular

$$a = \overline{y} - \frac{S_{XY}}{S_X^2}\overline{x} = 77.337 - \frac{29.386}{8.205^2} \cdot 155.207 = 77.337 - 0.436 \cdot 155.207 = 9.666$$
$$b = 0.436$$

De tal modo que si queremos calcular una predicción usando este modelo de regresión para el valor de ventas en febrero, conociendo que en enero hemos vendido $x = 172$ tendremos que substituir en nuestra recta $a + bx$, resultando la predicción

$$\widehat{y} = a + bx = 9.666 + 0.436 \cdot 172 = 84.658$$

3.2 Coeficiente de determinación R^2

El coeficiente de determinación se define como el cuadrado del coeficiente de correlación de pearson:

$$R^2 = \rho^2$$

Este valor da una idea **la precisión** con que podemos aproximar a los valores de Y usando la recta de regresión

El coeficiente de determinación está comprendido entre 0 y 1. Cuanto más se aproxime a 0, peor es el modelo de regresión lineal para describir la relación entre las variables. Cuanto más se aproxime a 1, mejor es el modelo.

No existe un criterio inequívoco sobre el mínimo valor exigible para que el modelo de regresión lineal sea aceptable. En general, se considera inadmisible un modelo con $R < 0.5$

3.3 El error estándar de la estimación

El **error estándar de la estimación** es una alternativa al coeficiente de determinación para estudiar la **la precisión** con que podemos aproximar a los valores de Y usando la recta de regresión.

Se define simplemente calculando un promedio los errores cometidos por la estimación en cada uno de los puntos de la muestra

$$S_e = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - a - bx_i)^2}{N-2}}$$

Donde

- (x_i, y_i) son los puntos de la muestra bidimensional
- a y b son los coeficientes de la recta de regresión de Y sobre X definidos anteriormente
- N es el número de datos

La idea es visualizando los datos en el diagrama de nube de puntos, cuanto más se peguen los datos a la recta menor será este error estandar de estimación,

3.3.1 Método sencillo para calcular S_e

Razonando sobre la fórmula anterior, en realidad podemos calcular S_e fácilmente usando

$$S_e = \sqrt{\frac{\sum_{i=1}^{N}(y_i - a - bx_i)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N} y_i^2 - a \cdot \sum_{i=1}^{N} y_i - b \cdot \sum_{i=1}^{N} x_i y_i}{N-2}}$$

Esto simplifica bastante el cálculo ya que calcular $\cdot \sum_{i=1}^{N} x_i y_i$, $\cdot \sum_{i=1}^{N} y_i$ y $\cdot \sum_{i=1}^{N} y_i^2$ suele hacerse como cálculo intermedio para calcular la covarianza y las varianzas

3.3.2 Ejemplo

En la siguente tabla x_i representan los años de antiguedad de los camiones de una empresa, e y_i representa los gastos de reparación.

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
5	7	35	25	49
3	7	21	9	49
3	6	18	9	36
1	4	4	1	16

Si calcularmos los coeficientes de la recta de regresión de Y sobre X obtendremos

$$a = 3.75$$
$$b = 0.75$$

De la tabla obtenemos $\sum xy = 78, \sum y = 24, \sum y^2 = 150$

luego

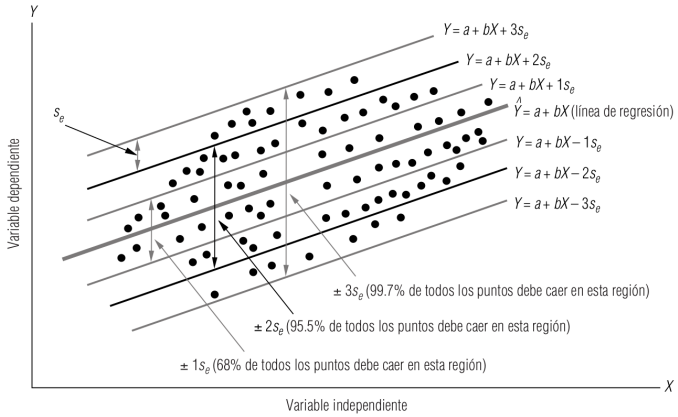
$$S_e = \sqrt{\frac{\sum_{i=1}^{N} y_i^2 - a \cdot \sum_{i=1}^{N} y_i - b \cdot \sum_{i=1}^{N} x_i y_i}{N-2}} = \frac{\sqrt{150 - 3.75 \cdot 24 - 0.75 \cdot 78}}{4-2} \cong 0.866$$

3.4 Interpretación del error estándar de la estimación

Como ocurría en el caso de la desviación estándar, mientras más grande sea el error estándar de la estimación, mayor será la dispersión de los puntos alrededor de la línea de regresión. De manera inversa, si $S_e = 0$, esperamos que la ecuación de estimación sea un estimador *perfecto* de la variable dependiente. En ese caso, todos los puntos caerían directamente sobre la línea de regresión y no habría puntos dispersos alrededor.

Suponiendo que los puntos observados siguen una distribución normal alrededor de la recta de regresión, podemos esperar encontrar el 68% de los puntos dentro de $\pm 1 S_e$ (o más menos 1 error estándar de la estimación), el 95.5% de los puntos dentro de $\pm 2 S_e$ y el 99.7% de los puntos dentro de $\pm 3 S_e$.

La siguiente figura resume adecuadamente este hecho



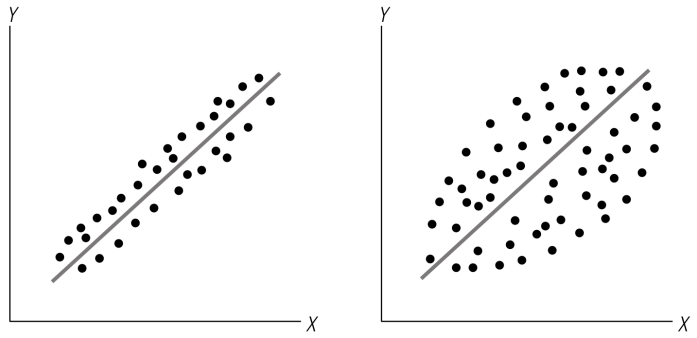
Es importante observar que en la figura anterior el error estándar de la estimación se mide a lo largo del eje Y, y no perpendicularmente desde la recta de regresión.

Siendo precisos, para que lo establecido en la figura sea correcto y los porcentajes comentados antes sean ciertos se debe asumir:

1. Los valores observados para Y tienen distribución normal alrededor de cada valor estimado de $\hat{Y}$.
2. La varianza de las distribuciones alrededor de cada valor posible de $\hat{Y}$ es la misma. Si esta segunda suposición no fuera cierta, entonces el error estándar en un punto de la recta de regresión podría diferir del error estándar en otro punto

3.5 Interpretación geométrica del error estándar de la estimación y R^2

Observemos las siguientes dos muestras bidimensionales



- La muestra de la izquierda está muy pegada a la recta de regresión, por lo que tendrá un error estandar S_e bajo, y un R^2 cercano a 1
- La muestra de la derecha está poco muy separada de la recta de regresión, por lo tanto será dificil de predecir usando regresión y tendrá un error estandar S_e alto, y un R^2 cercano a 0

3.6 Metodo completo del cálculo de la recta de regresión

Si tenemos unos datos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ... (x_N, y_N)$$

Para calcular la **recta de regresión de** Y **sobre** X (y por lo tanto predecir Y usando X debemos):

1. calcular $\overline{x}$, $\overline{y}$, S_{XY}, S_X y S_Y. Esto nos permitirá calcular el coeficiente de correlación de Pearson ρ

$$\rho = \frac{S_{XY}}{S_X S_Y}$$

2. Interpretar ρ: Un valor cercano a 1 es dependencia lineal directa. Un valor cercano a -1 indirecta, y un valor cercano a 0 implica no dependencia lineal.

3. Calcular los coeficientes a y b de la recta usando

$$a = \overline{y} - \frac{S_{XY}}{S_X^2}\overline{x}$$

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

4. Calcular el coeficiente de determinación $R^2 = \rho^2$ e interpretarlo. Un valor cercano a 0 significa que el modelo no aproxima bien, mientras que un valor cercano a 1 significa que el modelo es capaz de aproximar bien los valores de Y a partir de x. También podemos calcular el error estándar de estimación S_e.

Si ahora tenemos un nuevo valor de x y queremos estimar un valor de Y basta hacer

$$y = a + bx$$

con los valores obtenidos en el método anterior.

4 Ejercicios

1. Los siguientes datos corresponden al precio de un porducto y a la cantidad ofertada:

cantidad ofertada (miles) Y	precio (euros) X
1	3.5
5	5
10	8
15	8.5
20	12.5
25	13
30	15

   1. Calcula la covarianza y el coeficiente de correlación de Pearson e interprétalo. ¿Se cumple la ley de la oferta?
   2. Calcula la recta de regresión de Y sobre X
   3. Estima el valor de la cantidad ofertada que corresponderá al precio 18