ESTADISTICA DESCRIPTIVA BIDIMENSIONAL. REGRESION Y CORRELACION

1. Variable estadística bidimensional
2. Distribuciones de frecuencias
    a. Conjuntas
    b. Marginales
    c. Condicionadas
3. Representaciones gráficas
4. Medidas características de una Distribución de Frecuencias Bidimensional.
    a. Momentos respecto al origen
    b. Momentos respecto a las medias o centrales de orden (r, s)
5. Regresión Lineal. Correlación
    a. Introducción a los modelos de regresión
    b. El modelo de regresión lineal simple
    c. El coeficiente de correlación lineal
6. Otros modelos de regresión

1. Variable estadística bidimensional

Consideramos un conjunto de *dos características* para describir a cada uno de los individuos de la población. Representaremos por *(X, Y) la variable bidimensional* estudiada, donde X e Y son las variables unidimensionales correspondientes a la primera y segunda característica medidas para cada individuo.

Los dos caracteres observados no tienen por qué ser de la misma clase. Así, se nos pueden presentar las *situaciones* siguientes:

- Dos caracteres cualitativos
- Dos caracteres cuantitativos
- Uno cualitativo y otro cuantitativo.

En el caso de dos caracteres cuantitativos las *variables* pueden clasificarse:

- Dos discretas
- Dos continuas
- Una discreta y otra continua

2. Distribuciones de frecuencias

Distribuciones de frecuencias conjunta

Supongamos que estamos interesados en estudiar una muestra de tamaño n de una población en la que hemos considerado una variable estadística bidimensional (X, Y).

Llamamos **frecuencia absoluta del par** (x_i, x_j),que representaremos por n_{ij}, al número de individuos que presentan esa modalidad.

Llamamos **frecuencia relativa del par** (x_i, x_j), que representaremos por f_{ij}, al cociente entre la frecuencia absoluta y el tamaño de la muestra.

La distribución de frecuencias conjunta de la variable bidimensional (X, Y) es el resultado de organizar en una tabla las modalidades junto con las correspondientes frecuencias.

Ejemplo 1

Las calificaciones obtenidas por un grupo de diez alumnos en Filosofía (X) y en Literatura (Y) son:

X	3	4	6	7	5	8	7	3	5	4
Y	5	5	8	7	7	9	10	4	7	4

Ejemplo 2

La tabla siguiente corresponde a sesenta personas a las que se les midió y pesó:

X Y	[1.55, 1.65)	[1.65, 1.75)	[1.75, 1.80)
[50, 55)	2	1	0
[55, 60)	2	2	1
[60, 65)	1	3	2
[65, 70)	1	10	8
[70, 75)	4	5	5
[75, 80)	2	3	8

Distribuciones de frecuencias marginales

Llamaremos distribuciones marginales a las distribuciones de frecuencias unidimensionales de las variables X e Y, que se obtienen a partir de la distribución de frecuencias conjunta:

Ejemplo 3

A partir de la distribución conjunta del ejemplo 1, obtenemos las distribuciones marginales siguientes:

X	$n_{i.}$
3	2
4	2
5	2
6	1
7	2
8	1

Y	$n_{.j}$
4	2
5	2
7	3
8	1
9	1
10	1

Ejemplo 4

A partir de la distribución conjunta del ejemplo 2, obtenemos las distribuciones marginales siguientes:

X Y	[1.55, 1.65)	[1.65, 1.75)	[1.75, 1.80)	$n_{i.}$
[50, 55)	2	1	0	3
[55, 60)	2	2	1	5
[60, 65)	1	3	2	6
[65, 70)	1	10	8	19

[70, 75)	4	5	5	14
[75, 80)	2	3	8	13
$n._j$	12	24	24	60

De igual modo para las frecuencias relativas.

Distribuciones condicionadas

La distribución de X condicionada a $Y=y_j$, es la distribución unidimensional de X sabiendo que Y ha tomado la modalidad y_j.

Análogamente, se define la distribución de Y condicionada a $X=x_i$.

Ejemplo 5

A partir del ejemplo 2, la distribución de X condicionada a que Y=1.70, sería:

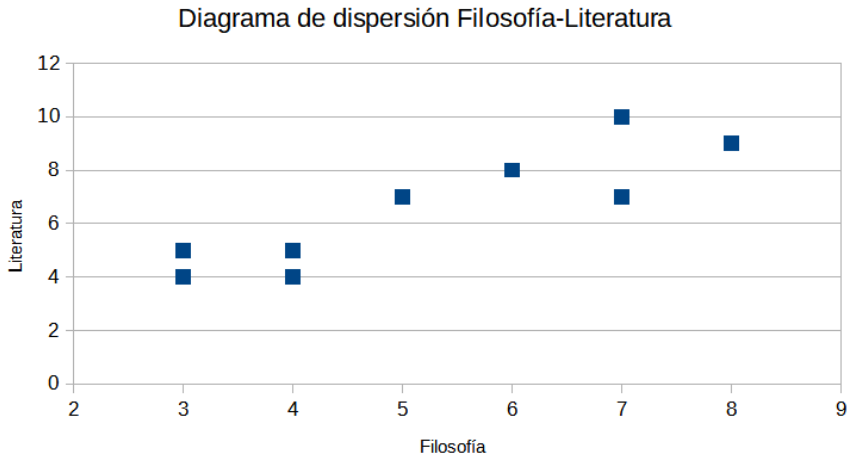
$X\|Y=1.70$	$n_{i\|2}$
[50, 55)	1
[55, 60)	2
[60, 65)	3
[65, 70)	10
[70, 75)	5
[75, 80)	3

3. Representaciones gráficas

La representación es más complicada que en el caso de las variables estadísticas unidimensionales. Cuando sean variables cuantitativas, utilizaremos los *diagramas de dispersión* y el *histograma tridimensional*.

Ejemplo 6

Para la distribución del ejemplo 1, el diagrama de dispersión sería el siguientes:



La representación gráfica se complica cuando las frecuencias son distintas de uno. En ese caso, recurrimos a un diagrama tridimensional.

4. Medidas características de una Distribución de Frecuencias Bidimensional

Para poder cuantificar el grado de relación que presentan dos variables, así como para establecer una recta que permita predecir una variable a partir de la otra, es

necesario establecer medidas o constantes para una distribución de frecuencias bidimensional.

Momentos

Momentos respecto al origen

La primera medida de relación entre dos variables que podemos pensar, es promediar su producto. Así surge el llamado **momento respecto al origen de orden (r, s),** cuya notación es $a_{r,s}$ y que se determina por

$$a_{r\,s} = \frac{1}{n}\sum_i \sum_j x_i^{\,r} \cdot y_j^{\,s} \cdot n_{i\,j}$$

Algunos momentos respecto al origen de interés:

- $a_{10} = \bar{x}$
- $a_{01} = \bar{y}$

Momentos respecto a las medias o centrales de orden (r, s).

Se definen mediante la siguiente expresión:

$$m_{r\,s} = \frac{1}{n}\sum_i \sum_j \left(x_i - \bar{x}\right)^r \cdot \left(y_j - \bar{y}\right)^s \cdot n_{i\,j}$$

Algunos momentos centrales de interés:

- $m_{20} = S_x^{\,2}$
- $m_{02} = S_y^{\,2}$

Especial interés tiene el momento:

$$m_{11} = \frac{1}{n}\sum_i \sum_j \left(x_i - \bar{x}\right) \cdot \left(y_j - \bar{y}\right) \cdot n_{i\,j}$$

Que recibe el nombre de **covarianza**, y se representa por $Cov(X,Y) = S_{xy}$.

Para su cálculo puede emplearse la siguiente igualdad:

$$S_{xy} = \frac{1}{n}\sum_i \sum_j x_i \cdot y_j \cdot n_{i\,j} - \bar{x} \cdot \bar{y}$$

Se denomina **vector de medias** al vector $\left(\bar{x}, \bar{y}\right)$.

Se denomina **matriz de varianzas-covarianzas** a la matriz:

$$S = \begin{pmatrix} S_x^{\,2} & S_{xy} \\ S_{xy} & S_y^{\,2} \end{pmatrix}$$

Al determinante de S se le llama **varianza generalizada.**

Ejemplo 7

Calcula las medias, varianzas y covarianza de la distribución bidimensional del ejemplo 1

Ejemplo 8

Calcula las medias, varianzas y covarianza de la distribución bidimensional del ejemplo 2

5. Regresión Lineal. Correlación

Es interesante estudiar la posible existencia de alguna relación de dependencia entre la variable X, **variable explicativa o independiente**, y la variable Y, de **respuesta o dependiente**; así como la construcción de algún modelo matemático que permita describir esa relación, caso de que exista.

Dos situaciones extremas que pueden presentarse son:

- Una **relación de dependencia exacta** entre las variables. Es el caso en el que X representa el radio de una circunferencia e Y su longitud. La relación existente es, $Y = 2\pi X$
- Las dos **variables son independientes**. Si X es la altura de una persona e Y su salario anual, no sabríamos qué valor de Y corresponde a un determinado valor de X.

Entre ambas, está aquella en que X contiene alguna información, aunque incompleta, de la variable Y. Podremos predecir un valor de Y, conocido un valor de X, mediante modelos de regresión. En este caso diremos que las variables X e Y son **dependientes**. Un ejemplo sería la estatura y el peso de una persona. Conocida la estatura podemos predecir su peso.

5.1. Introducción a los modelos de regresión

Distinguimos dos casos:

- **Modelos de regresión simple**
  Consideran una única variable explicativa X, intentando explicar el valor de la variable respuesta a partir de ella.
- **Modelos de regresión múltiple**
  Consideran dos o más variables explicativas.

En el caso de la regresión simple, se tratará de utilizar la información contenida en una muestra de n observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, de la variable bidimensional (X, Y) para construir modelos matemáticos de la forma

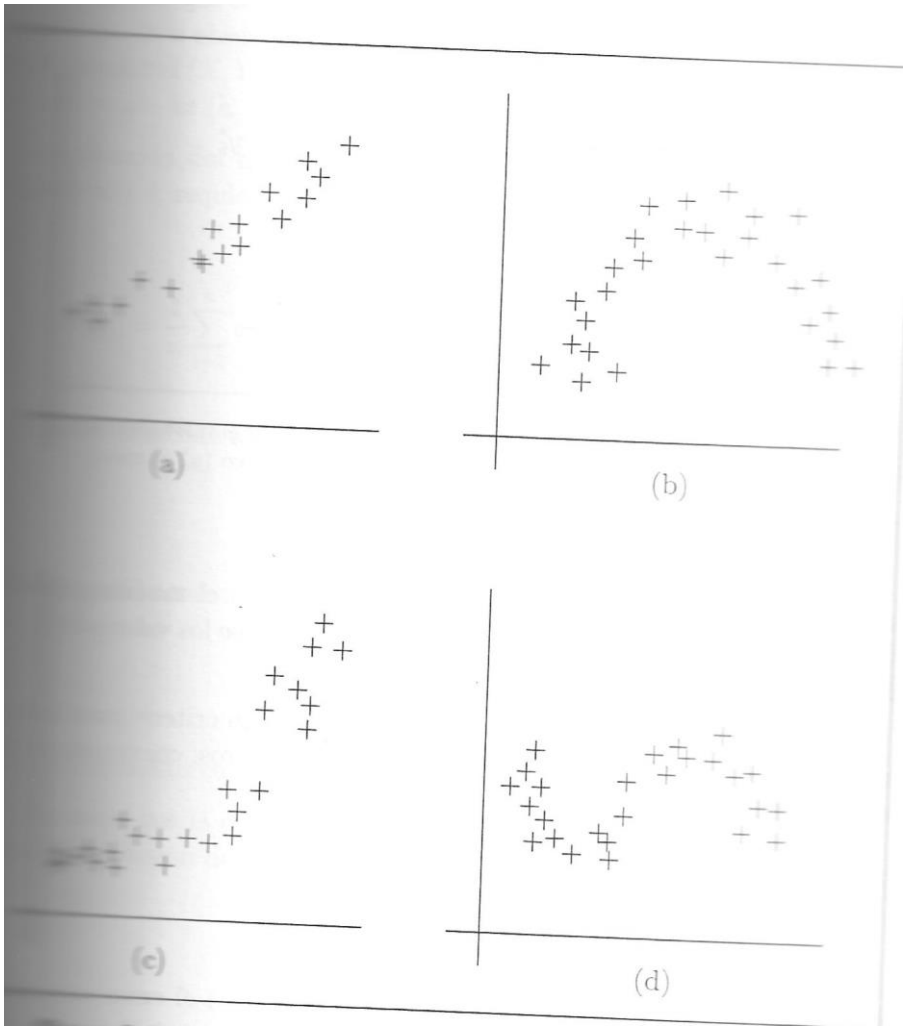
$$Y = m(X) + E$$

donde E representa el **error cometido** al predecir el valor de Y mediante el valor que resulta de evaluar la función m en un valor de X.

La **función de regresión**, m, determina la relación de dependencia entre las variables X e Y, pudiendo diferenciarse dos tipos:

- **Modelos paramétricos de regresión**: se supone que m tiene una forma predeterminada (lineal, $m(x) = a + bx$, parabólica, $m(x) = a + bx + cx^2, \ldots$)
- **Modelos no paramétricos de regresión:** se formulan condiciones muy generales acerca de la función m, debiendo determinarse ésta a partir de la información muestral.

Nos referiremos a los modelos paramétricos de regresión (principalmente al modelo de regresión lineal simple), lo cual supone la elección previa de la forma de la función de regresión m, que en la práctica puede resolverse observando el diagrama de dispersión.



Una vez elegida la forma del modelo de regresión, se tratará de determinar los valores de los parámetros desconocidos del modelo correspondientes a la curva que "mejor se ajusta"al diagrama de dispersión.

Para determinar los coeficientes, seguiremos el **método de los mínimos cuadrados.**

Método de los mínimos cuadrados

Sean $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, n pares de observaciones de una variable bidimensional (X, Y) y supongamos que se desea ajustar un modelo paramétrico de la forma

$$y = m(x, \theta_1, \theta_2, \ldots, \theta_k)$$

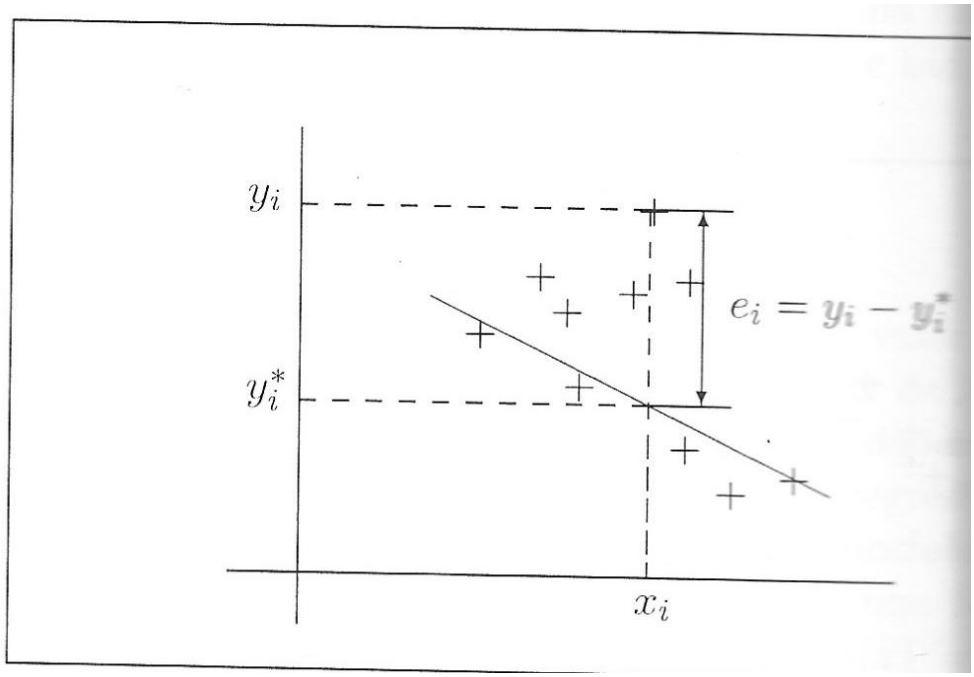
al diagrama de dispersión de (X, Y) donde $\theta_1, \theta_2, \ldots, \theta_k$ son las constantes desconocidas a determinar.

Por ejemplo, si $y = m(x, a, b) = a + bx$, los parámetros a determinar serían a y b.

El método de los mínimos cuadrados proporciona un criterio para la determinación de los valores desconocidos de los parámetros construido en base a la medida de ajuste

$M(\theta_1,\theta_2,\ldots,\theta_k)=\frac{1}{n}\sum_i {e_i}^2$; donde $e_i=y_i-m(x_i,\theta_1,\theta_2,\ldots,\theta_k)$

En el caso de un ajuste mediante la función $y=m(x,a,b)=a+bx$, gráficamente



En este caso e_i es el error cometido al sustituir el valor observado en la muestra y_i de la variable de respuesta Y, mediante el valor teórico y_i^* , que resulta de evaluar la función de regresión m en x_i.

De acuerdo con el método de mínimos cuadrados, los valores de los parámetros $\theta_1,\theta_2,\ldots,\theta_k$ del modelo que mejor se ajusta al diagrama de dispersión (X, Y) son los valores $\theta_1,\theta_2,\ldots,\theta_k$ que minimizan la función $M(\theta_1,\theta_2,\ldots,\theta_k)$.

5.2. El modelo de regresión lineal simple

Supongamos que los puntos del diagrama de dispersión obtenido para una muestra de n observaciones $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ de una variable aleatoria bidimensional (X, Y) tienden a alinearse en torno a una recta de ecuación $y=m(x)=a+bx$. Haremos el razonamiento para datos de frecuencia 1, siendo análogo para el caso general.

La aplicación del método de mínimos cuadrados para determinar los valores de a y de b, requiere minimizar la función

$$M(a,b)=\frac{1}{n}\sum_i {e_i}^2=\frac{1}{n}\sum_i\left[y_i-(a+bx_i)\right]^2=\frac{1}{n}\sum_i\left[y_i-a-bx_i\right]^2$$

Los posibles extremos son las soluciones del sistema

$$\begin{cases} \dfrac{\partial M}{\partial a}=0 \\ \dfrac{\partial M}{\partial b}=0 \end{cases} \Rightarrow \begin{cases} na+b\sum_i x_i=\sum_i y_i \\ a\cdot\sum_i x_i+b\cdot\sum_i x_i^{\,2}=\sum_i x_i\cdot y_i \end{cases} \Rightarrow \begin{cases} a=\bar{y}-\dfrac{S_{xy}}{S_x^{\,2}}\cdot\bar{x} \\ b=\dfrac{S_{xy}}{S_x^{\,2}} \end{cases}$$

De manera que la recta de regresión ajustada al diagrama de dispersión de (X, Y) mediante el método de los mínimos cuadrados es

$$y=m(x)=a+bx$$

O bien, sustituyendo por sus valores

$$y-\bar{y}=\frac{S_{xy}}{S_x^{\,2}}(x-\bar{x})$$

Que denominaremos **recta de regresión de Y sobre X**, ya que ha sido obtenida considerando que Y es la variable respuesta y que X es la variable explicativa.

Sin embargo, si intercambiamos los papeles, de X e Y, obtenemos una recta de regresión distinta

$$x-\bar{x}=\frac{S_{xy}}{S_y^{\,2}}(y-\bar{y})$$

Que denominaremos **recta de regresión de X sobre Y**.

Llamaremos **coeficientes de regresión** a las pendientes de las rectas de regresión de Y sobre X, $\beta_{yx}=\dfrac{S_{xy}}{S_x^{\,2}}$, y de X sobre Y, $\beta_{xy}=\dfrac{S_{xy}}{S_y^{\,2}}$.

Si las rectas de regresión coinciden (dependencia lineal exacta), entonces

$$\beta_{yx}=\left(\beta_{xy}\right)^{-1}\Rightarrow \beta_{yx}\cdot\beta_{xy}=1$$

Predicción con el modelo de regresión lineal simple

El modelo de predicción es un problema de fácil solución: el valor de la predicción será el resultado de evaluar la función de regresión ajustada en el valor particular de la variable o variables explicativas.

Así, en el supuesto de que el modelo ajustado corresponde a la recta de regresión de Y sobre x, la predicción del valor de Y, conocido un valor particular x_0 de X, se obtiene sustituyendo en la ecuación del modelo ajustado x por el valor numérico x_0 , de manera que la predicción de Y con el modelo de regresión simple es $y_0=a+bx_0$.

Debemos esperar *resultados razonables* cuando el valor x_0 pertenezca al intervalo determinado por el mínimo y máximo valor de los valores x_i observados en la muestra. Fuera de ese rango, al no tener información, pudiera no ser apropiado utilizar el modelo de regresión ajustado.

Adecuación del modelo de regresión lineal simple: la razón de correlación

Ahora pretendemos conocer si el modelo ajustado es adecuado para describir la relación de dependencia entre las variables X e Y de manera que podamos

utilizarlo, por ejemplo, para obtener predicciones razonablemente válidas de Y a partir de valores particulares de X.

Una vez obtenida la recta de regresión $y = a + bx_0$, para cada valor x_i de X medido en la muestra (i : 1, 2,…, n) podemos medir para Y los valores observados y_i y las predicciones $y_i = a + bx_i$, de manera que el error (llamado residuo) e_i que cometemos al predecir el valor de Y en $x=x_i$, con el modelo de la recta de regresión $e_i = y_i - y_i$.

Por tanto, la **varianza no explicada**

$$S_R{}^2 = \frac{1}{n}\sum_i e_i{}^2 = \frac{1}{n}\sum_i \left(y_i - y_i\right)^2 = \frac{1}{n}\sum_i \left(y_i - a - bx_i\right)^2$$

que representa **la parte de la variación de Y** que no es capaz de explicar la recta de regresión, puede interpretarse como una medida de la bondad de ajuste del modelo de regresión lineal: valores grandes de la varianza no explicada indican que el modelo es poco adecuado y valores pequeños de la varianza no explicada indican que el modelo puede ser adecuado.

Sin embargo, $S_R{}^2$ tiene un inconveniente importante que impide que pueda utilizarse para juzgar la bondad de ajuste de la recta de regresión y es que depende de las unidades de medida. Una forma de evitar este problema es dividir la varianza no explicada para la varianza total de Y, $S_y{}^2$, de manera que utilizaremos el cociente adimensional $\frac{S_R{}^2}{S_y{}^2}$ que representa la proporción de variación de Y no explicada por la recta de regresión.

No obstante, en la práctica, la medida que suele utilizarse para juzgar la bondad de ajuste de la recta de regresión es la **razón de correlación**, también llamado **coeficiente de determinación**

$$R^2 = 1 - \frac{S_R{}^2}{S_y{}^2} = \frac{S_y{}^2 - S_R{}^2}{S_y{}^2}$$

que representa la proporción de la variación de Y explicada por el modelo de la regresión. Este coeficiente puede calcularse para cualquier medida de regresión.

En el caso particular del modelo de regresión lineal, se obtiene

$$R^2 = \frac{S_{xy}{}^2}{S_x{}^2 \cdot S_y{}^2} = \beta_{xy} \cdot \beta_{yx}$$

En cuanto a la interpretación de esta medida, es claro que valores de R^2 cercanos a uno indican un buen ajuste del modelo, mientras que valores cercanos al cero indican un mal ajuste del modelo. En la práctica, podemos convenir en aceptar el modelo de regresión ajustado si R^2 es mayor que 0.9, esto es, si explica el 90% de la variación de Y.

5.3. El coeficiente de correlación lineal

El propósito del estudio de la correlación es la construcción de medidas del grado de dependencia o asociación entre variables estadísticas, siendo la más popular de estas el **coeficiente de correlación lineal de X e Y** como el número

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

que es una medida del grado de dependencia lineal entre las variables X e Y.

Propiedades

- El coeficiente de correlación es una medida adimensional.
- En el modelo de regresión lineal, $R^2 = r_{xy}$, en consecuencia, como $0 \le R^2 \le 1$, se verifica que $-1 \le r_{xy} \le 1$. Pudiendo juzgarse la bondad de ajuste del modelo de regresión lineal en función del valor calculado para el coeficiente de regresión lineal.

  Entonces
  - Si $r_{xy} = \pm 1$, existe **dependencia lineal exacta** entre X e Y. Coinciden las rectas de regresión**.**
  - Si $r_{xy} = 0$, no existe dependencia lineal entre X e Y, lo que no significa necesariamente que X e Y sean independientes. Diremos que son **incorreladas**.
  - Como $r_{xy} = r_{yx}$ y las pendientes de las rectas de regresión son

    $$\beta_{yx} = r_{xy} \cdot \frac{S_y}{S_x} \; ; \; \beta_{xy} = r_{xy} \cdot \frac{S_x}{S_y}$$

    de manera que si $r_{xy} > 0$ la recta de regresión es creciente y si $r_{xy} < 0$, la recta de regresión es decreciente. Además, $\beta_{xy} \cdot \beta_{yx} = {r_{xy}}^2$

A partir de la definición anterior del coeficiente de correlación lineal, se define la **matriz de correlaciones** de (X, Y) como la matriz simétrica

$$C = \begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}$$

Definiéndose, en general, la matriz de correlaciones de una variable k-dimensional (X_1, X_2, …, X_k) como la matriz simétrica $C = \left(r_{ij}\right)$, siendo $r_{ij}, \left(1 \le i \le k, 1 \le j \le k\right)$ el coeficiente de correlación lineal entre X_i y X_j.

6. Otros modelos de regresión

Destacamos los siguientes:

El modelo potencial o multiplicativo

Se trata de ajustar al diagrama de dispersión la función de regresión $y = m(x) = ax^b$, pudiéndose reducir (en el supuesto de que X e Y tomen valores positivos) al caso de la regresión lineal simple mediante la transformación logarítmica,

$$\log y = \log a + b \log x \, .$$

El modelo recíproco o inverso

La función de regresión a ajustar al diagrama de dispersión es $y = m(x) = \frac{1}{a + bx}$, que (supuesto que Y es no nula) puede reducirse al caso de la regresión lineal simple mediante la transformación inversa,

$$\frac{1}{y} = a + bx \, .$$

El modelo exponencial

Se trata de ajustar al diagrama de dispersión la función de regresión $y = m(x) = ab^x$, que (supuesto Y toma valores positivos) puede reducirse también al caso de la regresión lineal simple mediante la transformación logarítmica

$$\log y = \log a + x \log b \, .$$

El modelo de regresión polinómica

Este modelo generaliza el caso de la regresión lineal simple, siendo la función de regresión a ajustar al diagrama de dispersión un polinomio de grado p,

$$y = m(x) = \theta_0 + \theta_1 x + \cdots + \theta_p x^p$$

El modelo de regresión lineal múltiple

Es la generalización del modelo de regresión lineal simple al caso en que se disponga de k variables explicativas X_1, X_2, …, X_k, tratándose de ajustar al diagrama de dispersión, ((k+1)-dimensional), la función de regresión

$$y = m(x_1, x_2, \ldots, x_k) = \theta_0 + \theta_1 x + \cdots + \theta_k x^k$$