TEMAS DE MATEMÁTICAS
(Oposiciones de Secundaria)

TEMA 62

SERIES ESTADÍSTICAS BIDIMENSIONALES. COEFICIENTE DE VARIACIÓN. VARIABLE NORMALIZADA. APLICACIÓN AL ANÁLISIS, INTERPRETACIÓN Y COMPARACIÓN DE DATOS ESTADÍSTICOS.

1. Distribuciones Bidimensionales de Frecuencias.
   1.1. Independencia y Relación Funcional de dos Variables.
   1.2. Tablas de Correlación y de Contingencia.
   1.3. Distribuciones Marginales.
   1.4. Distribuciones Condicionadas.
   1.5. Independencia Estadística.
2. Representaciones Gráficas.
3. Momentos de Distribuciones Bidimensionales.
   3.1. Momentos Respecto al Origen.
   3.2. Momentos Respecto a las Medias.
   3.3. Cálculo de los Momentos Centrales en función de los Momentos Respecto al Origen.
   3.4. Método Reducido para el Cálculo de Varianza y Covarianza.
   3.5. Valor de la Covarianza en caso de Independencia Estadística.
4. Ajuste.
   4.1. Método de los Mínimos Cuadrados.
      4.1.1. Ajuste de una Recta.
      4.1.2. Ajuste de una Parábola.
      4.1.3. Ajuste Hiperbólico.
      4.1.4. Ajuste Potencial.
      4.1.5. Ajuste Exponencial.
   4.2. Método de los Momentos.
5. Regresión.
   5.1. Regresión Lineal.
      5.1.1. Recta de Regresión de Y sobre X.
      5.1.2. Recta de Regresión de X sobre Y.
   5.2. Coeficientes de Regresión.
6. Correlación.
   6.1. Campo de Variación de R y su Interpretación.
   6.2. Coeficiente de Correlación Lineal.
   6.3. Interpretación Analítica de r.
   6.4. Interpretación Geométrica de r.
7. Varianza debida a la Regresión y Coeficiente de Determinación Lineal.
8. Aplicaciones de la Regresión y la Correlación.
   8.1. Uso y Abuso de la Regresión.
   8.2. Predicción.

Bibliografía Recomendada.

TEMA 62

SERIES ESTADÍSTICAS BIDIMENSIONALES. COEFICIENTE DE VARIACIÓN. VARIABLE NORMALIZADA. APLICACIÓN AL ANÁLISIS, INTERPRETACIÓN Y COMPARACIÓN DE DATOS ESTADÍSTICOS.

1. DISTRIBUCIONES BIDIMENSIONALES DE FRECUENCIAS.

Si estudiamos sobre la misma población dos caracteres cuantitativos X e Y y los medimos en las mismas unidades estadísticas, obtenemos dos series estadísticas de las variables X e Y. Considerando simultáneamente ambas series, el par de valores (x_i, y_i) le corresponde una variable estadística Bidimensional.

Es posible estudiar de forma separada la distribución de la población según el carácter X o Y, obteniendo $\overline{x}, S_x, \overline{y}, S_y$ o cualquier otro parámetro. Pero puede ser interesante considerar de forma simultánea los dos caracteres, con el objetivo de determinar las posibles relaciones entre ellos y así poder responder a preguntas como ¿Existe algún tipo de relación entre los caracteres X e Y?.

Vamos a ver instrumentos estadísticos que nos van a permitir obtener la existencia o no de coincidencias entre los valores de dos variables y, a partir de esas coincidencias, formular la hipótesis de una relación causal entre los dos caracteres.

Si existen coincidencias estadísticas entre los valores de dos caracteres, o lo que es lo mismo, si existe relación entre las dos variables, las coincidencias pueden ser más o menos fuertes, y la intensidad de la relación puede variar entre ausencia total de relación o ligazón perfecta.

1.1. Independencia y Relación Funcional de dos Variables.

DEF Diremos que dos variables son independientes cuando no existe relación entre ambas. Inversamente, cuando la relación entre dos variables es perfecta, diremos que están relacionadas funcionalmente, lo cual implica que su relación puede expresarse como y=f(x).

DEF Diremos que Y depende funcionalmente de X cuando podamos establecer una aplicación que nos transforme los elementos de X en elementos de Y.

Desde el punto de vista de la Estadística, lo que realmente nos interesa es que podemos determinar los elementos de Y conocidos los de X, o viceversa. Pero esa circunstancia no será muy habitual. Existen características como la estatura y el peso, consumo y renta, etc. en los que aun existiendo interrelación, es imposible definir una aplicación en el sentido estrictamente matemático. Es decir, no dependen funcionalmente una de otra.

Estadísticamente hablando, es claro que el peso depende en cierta forma de la estatura, el consumo depende de la renta, etc. Este tipo de relación no expresable a través de una determinada aplicación es la conocida como Dependencia Estadística. Y

este tipo de dependencia si admite grados, ya que puede haber dependencias más o menos fuertes.

Estos tipos de dependencia se pueden expresar gráficamente mediante un segmento de la recta real, donde en un extremo situamos la dependencia funcional y en el otro la independencia. Los puntos intermedios del segmento se corresponden con los diferentes grados de dependencia estadística.

1.2. Tablas de Correlación y de Contingencia.

Dada una población, en la que estudiamos simultáneamente dos caracteres X e Y, podemos representar la distribución mediante ternas de la forma (x_i, y_j, n_{ij}), donde x_i e y_j son dos valores cualesquiera y n_{ij} es la frecuencia absoluta conjunta del valor i-ésimo de X y j-ésimo de Y.

Los resultados se pueden representar en una tabla de doble entrada conocida como Tabla de Correlación.

X \ Y	y_1	y_2	…	y_n	$\mathbf{n_{i\cdot}}$
x_1	n_{11}	n_{12}	…	n_{1n}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	…	n_{2n}	$n_{2\cdot}$
…	…	…	…	…	…
x_m	n_{m1}	n_{m2}	…	n_{mn}	$n_{m\cdot}$
$\mathbf{n_{\cdot j}}$	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot n}$	N

Si la distribución bidimensional es de atributos, la tabla de doble entrada recibe el nombre de Tabla de Contingencia.

1.3. Distribuciones Marginales.

A partir de una distribución bidimensional podemos realizar el estudio de cada una de las variables de forma aislada. Tendríamos así dos distribuciones unidimensionales las cuales serían las correspondientes a X e Y respectivamente.

Para poder obtenerlas, necesitamos determinar las frecuencias marginales. La distribución marginal de X se halla obteniendo cuantas veces se repite el valor x_i, independientemente de que aparezca conjuntamente o no con algún valor de Y. Así, tenemos que la distribución marginal de X sería:

X	$n_{i\cdot}$
x_1	$n_{1\cdot} = \sum_{j=1}^{n} n_{1j}$
x_2	$n_{2\cdot} = \sum_{j=1}^{n} n_{2j}$
…	…
x_m	$n_{m\cdot} = \sum_{j=1}^{n} n_{mj}$

Análogamente obtendríamos la distribución marginal de Y.

1.4. Distribuciones Condicionadas.

Se pueden formar otro tipo de distribuciones unidimensionales en las que previamente haría falta definir una condición. En general, las distribuciones de X condicionadas a que Y tome un determinado valor (por ejemplo y_j) son:

x_i/y_j	$n_{i/j}$
x_1	n_{1j}
x_2	n_{2j}
…	…
x_m	n_{mj}
	$n_{\cdot j}$

De forma análoga construiríamos las distribuciones de Y condicionadas a que X tome un determinado valor.

La frecuencia relativa de la distribución condicionada a algún valor de y es:

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}}$$

Análogamente, la frecuencia relativa de la distribución condicionada a algún valor de x es:

$$f_{j/i} = \frac{n_{ij}}{n_{i\cdot}}$$

1.5. Independencia Estadística.

DEF Diremos que dos variables X e Y son independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales. Es decir:

$$\frac{n_{ij}}{N} = \frac{n_{i\cdot}}{N}\cdot\frac{n_{\cdot j}}{N} \quad \forall i, j$$

En este caso, las frecuencias relativas condicionadas serán:

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}\dfrac{n_{\cdot j}}{N}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \qquad f_{j/i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{i\cdot}\dfrac{n_{\cdot j}}{N}}{n_{i\cdot}} = \frac{n_{\cdot j}}{N}$$

Como vemos, las frecuencias relativas condicionadas son iguales a sus correspondientes frecuencias relativas marginales, lo que nos indica que el condicionamiento no existe. Las variables son independientes, puesto que en las

distribuciones marginales se estudia el comportamiento de una variable con independencia de los valores que pueda tomar la otra.

2. REPRESENTACIONES GRÁFICAS.

La representación gráfica más utilizada consiste en representar cada pareja de valores mediante un punto en un sistema de ejes coordenados. Por tanto, la distribución vendrá dada por un conjunto de puntos que recibe el nombre de Nube de Puntos o Diagrama de Dispersión. Cuando una pareja de valores está repetida, junto a la representación del punto correspondiente se indica el valor de su frecuencia.

La representación gráfica de la nube de puntos puede hacerse tanto con datos agrupados (las marcas de clase son las que se representan) como con datos sin agrupar. En el diagrama de tres dimensiones y utilizando los límites de intervalos (no las marcas de clase), el "escalograma" más adecuado es el constituido por paralelepípedos cuyo volumen sea la correspondiente frecuencia, y los lados de la base cada una de las amplitudes de los respectivos intervalos de las variables, y donde n_{ij} es el volumen del paralelepípedo y h_{ij} la altura del mismo.

$$n_{ij} = (L_i - L_{i-1}) \cdot (L_j - L_{j-1}) \cdot h_{ij}$$

3. MOMENTOS DE DISTRIBUCIONES BIDIMENSIONALES.

Al igual que se definen los momentos en las distribuciones unidimensionales, podemos hacerlo en las bidimensionales. Por tanto, podemos distinguir entre momentos respecto al origen y momentos respecto a la media.

3.1. Momentos Respecto al Origen.

DEF Definimos el momento de orden r,s respecto al origen para la distribución (x_i, y_j, n_{ij}) como

$$\alpha_{rs} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i^r y_j^s \cdot \frac{n_{ij}}{N}$$

Podemos calcular los momentos de primer orden:

$$\alpha_{10} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i^1 y_j^0 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m} x_i \cdot \sum_{j=1}^{n} \frac{n_{ij}}{N} = \sum_{i=1}^{m} x_i \cdot \frac{n_{i\cdot}}{N} = \bar{x}$$

$$\alpha_{01} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i^0 y_j^1 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m} \sum_{j=1}^{n} y_j \cdot \frac{n_{ij}}{N} = \sum_{j=1}^{n} y_j \cdot \sum_{i=1}^{m} \frac{n_{ij}}{N} = \sum_{j=1}^{n} y_j \cdot \frac{n_{\cdot j}}{N} = \bar{y}$$

También resulta sencillo calcular los momentos de segundo orden:

$$a_{20} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i^2 y_j^0 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i^2 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m} x_i^2 \cdot \sum_{j=1}^{n}\frac{n_{ij}}{N} = \sum_{i=1}^{m} x_i^2 \cdot \frac{n_{i\cdot}}{N}$$

$$a_{02} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i^0 y_j^2 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} y_j^2 \cdot \frac{n_{ij}}{N} = \sum_{j=1}^{n} y_j^2 \cdot \sum_{i=1}^{m}\frac{n_{ij}}{N} = \sum_{j=1}^{n} y_j^2 \cdot \frac{n_{\cdot j}}{N}$$

$$a_{01} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i\ y_j \cdot \frac{n_{ij}}{N}$$

3.2. Momentos Respecto a las Medias.

DEF Definimos el momento de orden r,s respecto a las medias para la distribución (x_i, y_j, n_{ij}) como:

$$m_{rs} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^r \cdot \left(y_j - \bar{y}\right)^s \cdot \frac{n_{ij}}{N}$$

Los momentos de primer orden son

$$m_{10} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^1 \left(y_j - \bar{y}\right)^0 \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)\ \frac{n_{ij}}{N} = \sum_{i=1}^{m} \left(x_i - \bar{x}\right)\ \sum_{j=1}^{n}\frac{n_{ij}}{N} = \sum_{i=1}^{m} \left(x_i - \bar{x}\right)\ \frac{n_{i\cdot}}{N} = 0$$

$$m_{01} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^0 \left(y_j - \bar{y}\right)^1 \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(y_j - \bar{y}\right)\ \frac{n_{ij}}{N} = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)\ \sum_{i=1}^{m}\frac{n_{ij}}{N} = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)\ \frac{n_{\cdot j}}{N} = 0$$

Los Momentos de segundo orden son:

$$m_{20} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^2 \left(y_j - \bar{y}\right)^0 \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^2 \frac{n_{ij}}{N} = \sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2 \sum_{j=1}^{n}\frac{n_{ij}}{N} = \sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2\ \frac{n_{i\cdot}}{N} = S_X^2$$

$$m_{02} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^0 \left(y_j - \bar{y}\right)^2 \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2 \frac{n_{ij}}{N} = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2 \sum_{i=1}^{m}\frac{n_{ij}}{N} = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2\ \frac{n_{\cdot j}}{N} = S_Y^2$$

$$m_{11} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^1 \cdot \left(y_j - \bar{y}\right)^1 \cdot \frac{n_{ij}}{N} = S_{XY}$$

DEF Llamamos Covarianza al momento μ_{11}, que también se representa por S_{XY}.

3.3. Cálculo de los Momentos Centrales en Función de los Momentos respecto al Origen.

Al igual que sucede en las distribuciones unidimensionales, los momentos centrales de una distribución bidimensional pueden expresarse en función de los momentos respecto del origen.

Veamos:

$$m_{20} = \sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2 \frac{n_{i\cdot}}{N} = \sum_{i=1}^{m} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\frac{n_{i\cdot}}{N} = \sum_{i=1}^{m} x_i^2 \frac{n_{i\cdot}}{N} - 2\bar{x}\sum_{i=1}^{m} x_i \frac{n_{i\cdot}}{N} + \bar{x}^2 \sum_{i=1}^{m} \frac{n_{i\cdot}}{N} =$$

$$= \alpha_{20} - 2\bar{x}\alpha_{10} + \bar{x}^2 = \alpha_{20} - 2\alpha_{10}^2 + \alpha_{10}^2 = \alpha_{20} - \alpha_{10}^2$$

Por tanto tenemos que $S_X^2 = m_{20} = \alpha_{20} - \alpha_{10}^2$

De forma análoga comprobaríamos que $S_Y^2 = m_{02} = \alpha_{02} - \alpha_{01}^2$

Además, de la covarianza podemos decir:

$$S_{XY} = m_{11} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^1 \cdot \left(y_j - \bar{y}\right)^1 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(x_i y_j - \bar{x}y_j - \bar{y}x_i + \bar{x}\bar{y}\right) \cdot \frac{n_{ij}}{N} =$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j \cdot \frac{n_{ij}}{N} - \bar{x}\sum_{i=1}^{m}\sum_{j=1}^{n} y_j \cdot \frac{n_{ij}}{N} - \bar{y}\sum_{i=1}^{m}\sum_{j=1}^{n} x_i \cdot \frac{n_{ij}}{N} + \bar{x}\bar{y}\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{n_{ij}}{N} =$$

$$= \alpha_{11} - \bar{x}\alpha_{01} - \bar{y}\alpha_{10} + \bar{x}\bar{y} = \alpha_{11} - \alpha_{10}\alpha_{01} - \alpha_{01}\alpha_{10} + \alpha_{10}\alpha_{01} = \alpha_{11} - \alpha_{10}\alpha_{01}$$

Nos queda que la covarianza es $S_{XY} = m_{11} = \alpha_{11} - \alpha_{10}\alpha_{01}$

3.4. Método Reducido para el Cálculo de Varianza y Covarianza.

En aquellos casos en los que nos pueda parecer conveniente, podemos realizar determinados cambios de variable para así simplificar los cálculos.

Los cambios de variable siempre serán los mismos:

$$x_i' = \frac{x_i - O_1}{c_1} \qquad y_j' = \frac{y_j - O_2}{c_2}$$

siendo O_1 y O_2 orígenes de trabajo arbitrarios que se procuran sean puntos centrales de la distribución.

Así, sabemos que:

$$\bar{x} = c_1\bar{x}' + O_1$$
$$\bar{y} = c_2\bar{y}' + O_2$$
$$S_X^2 = c_1^2 (S_X')^2$$
$$S_Y^2 = c_2^2 (S_Y')^2$$
$$S_{XY} = c_1 c_2 S_{XY}'$$

3.5. Valor de la Covarianza en caso de Independencia Estadística.

Según hemos visto, la covarianza se podía expresar como $S_{XY} = m_{11} = a_{11} - a_{10}a_{01}$

La condición de independencia estadística era $\frac{n_{ij}}{N} = \frac{n_{i\cdot}}{N}\cdot\frac{n_{\cdot j}}{N} \quad \forall i, j$

Calculemos, según esta condición, el valor de α_{11}

$$a_{11} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i^1 y_j^1 \cdot \frac{n_{ij}}{N} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i\, y_j \cdot \frac{n_{i\cdot}}{N}\cdot\frac{n_{\cdot j}}{N} = \sum_{i=1}^{m} x_i \frac{n_{i\cdot}}{N}\cdot\sum_{j=1}^{n} y_j \cdot \frac{n_{\cdot j}}{N} = a_{10}\cdot a_{01}$$

Luego, cuando las variables son independientes, la covarianza es nula.

En cambio el recíproco no tiene por qué ser cierto.

4. AJUSTE.

Sea (x_i, y_j, n_{ij}) una distribución bidimensional en la que suponemos que existe relación entre las variables aleatorias X e Y. Si representamos en un sistema de ejes coordenados los pares de valores de ambas variables, el problema del ajuste consiste en obtener la ecuación de una curva que pase cerca de los puntos y se adapte lo mejor posible a los mismos, cumpliendo unas determinadas condiciones.

Cuando pretendemos realizar un ajuste nos encontramos con dos problemas:

1) Elegir el mejor tipo de curva que se adapte a los datos disponibles, es decir, aquella que mejor represente la relación existente entre X e Y. Es importante, sólo a modo de orientación, ver la representación gráfica de los puntos.
2) Fijado el tipo de curva a través de su ecuación en forma explícita con un cierto número de parámetro, determinar éstos mediante las condiciones que se impongan según el procedimiento de ajuste planteado.

4.1. Métodos de los Mínimos Cuadrados.

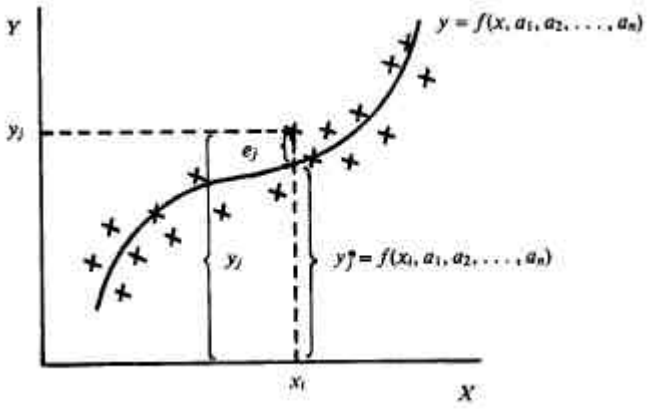
Dados los puntos (x_1,y_1), (x_2,y_2), ..., (x_m,y_m), podemos elegir una función de ajuste definida por:

$$y = f(x,a_1,a_2,...,a_n)$$

en la que intervienen n parámetros $(a_1,a_2,...,a_n)$ con $n<m$.

Para cada valor x_i de la variable X, tenemos dos valores para Y:

1) El valor, y_j, observado en la propia distribución.
2) El valor, que denotaremos por y_j^*, que se obtiene de sustituir x por x_i en la función f.



Para cada término x_i, existe una diferencia entre el valor observado en la distribución y el obtenido en forma teórica, que llamaremos Residuo.

$$e_j = y_j - y_j^*$$

El método de mínimos cuadrados consiste en determinar los parámetros $a_1, a_2, ..., a_n$ de tal forma que los residuos sean mínimos.

Si tomamos la suma de todos los residuos

$$\sum_i \sum_j \left(y_j - y_j^*\right) n_{ij}$$

nos encontramos con dos problemas a la hora de minimizar la expresión. El primero es que tenemos residuos de diferente signo, los cuales se compensarán en la suma, pudiendo dar una suma muy pequeña para residuos muy grandes. El segundo problema es que la determinación de los parámetros no es única, ya que obtendríamos diferentes conjuntos de parámetros que arrojarían la misma suma mínima de los residuos.

Para solucionar estos problemas siguiendo con el método de mínimos cuadrados, lo que vamos a hacer es tratar de minimizar la expresión:

$$f = \sum_i \sum_j \left(y_j - y_j^*\right)^2 n_{ij}$$

Como los valores teóricos son los obtenidos a partir de la curva ajustada, tenemos que la expresión a minimizar se queda como:

$$f = \sum_i \sum_j \left(y_j - f(x_i, a_1, a_2, \ldots, a_n)\right)^2 n_{ij}$$

para lo cual, la condición necesaria es que las primeras derivadas parciales respecto a cada uno de los parámetros se anulen.

El sistema que se obtiene recibe el nombre de Sistema de Ecuaciones Normales.

$$\left.\begin{aligned}
\frac{\partial f}{\partial a_1} &= 2\sum_i\sum_j \left[y_j - f(x_i; a_1, a_2, \ldots, a_n)\right]\cdot n_{ij}\cdot(-f'_{a_1}) = 0\\
\frac{\partial f}{\partial a_2} &= 2\sum_i\sum_j \left[y_j - f(x_i; a_1, a_2, \ldots, a_n)\right]\cdot n_{ij}\cdot(-f'_{a_2}) = 0\\
&\cdots\\
\frac{\partial f}{\partial a_n} &= 2\sum_i\sum_j \left[y_j - f(x_i; a_1, a_2, \ldots, a_n)\right]\cdot n_{ij}\cdot(-f'_{a_n}) = 0
\end{aligned}\right\}$$

Resolviendo este sistema determinamos los valores de los parámetros, así como la propia función f.

A continuación vamos a utilizar el método descrito para ajustar algunas funciones que corrientemente se suelen presentar.

4.1.1. Ajuste de una Recta.

Dada una nube de puntos, vamos a tratar de ajustarla mediante una recta de ecuación

$$y^*_j = a + bx_i$$

Para determinar los coeficientes a y b buscaremos el mínimo de la función:

$$f = \sum_i\sum_j \left(y_j - y^*_j\right)^2 n_{ij} = \sum_i\sum_j \left(y_j - (a + bx_i)\right)^2 n_{ij} = \sum_i\sum_j \left(y_j - a - bx_i\right)^2 n_{ij}$$

para lo cual, obtendremos las derivadas parciales de la función con respecto a los parámetros. Al igualar a cero ambas expresiones, resolvemos el sistema de dos ecuaciones que aparece.

$$\left.\begin{aligned}
\frac{\partial f}{\partial a} &= 2\sum_i\sum_j \left(y_j - a - bx_i\right)(-1)n_{ij} = 0\\
\frac{\partial f}{\partial b} &= 2\sum_i\sum_j \left(y_j - a - bx_i\right)(-x_i)n_{ij} = 0
\end{aligned}\right\}$$

Dividiendo ambos miembros por –2, nos queda:

$$\left.\begin{aligned}
\sum_i\sum_j \left(y_j - a - bx_i\right)n_{ij} &= 0\\
\sum_i\sum_j \left(y_j - a - bx_i\right)(x_i)n_{ij} &= 0
\end{aligned}\right\}$$

Operando y cambiando términos de un miembro a otro:

$$\left.\begin{aligned}\sum_i\sum_j y_j n_{ij} &= a\sum_i\sum_j n_{ij} + b\sum_i\sum_j x_i n_{ij}\\ \sum_i\sum_j y_j x_i n_{ij} &= a\sum_i\sum_j x_i n_{ij} + b\sum_i\sum_j x_i^2 n_{ij}\end{aligned}\right\}$$

Podemos resumir la expresión anterior en:

$$\left.\begin{aligned}\sum_j y_j n_{\cdot j} &= aN + b\sum_i x_i n_{i\cdot}\\ \sum_i\sum_j y_j x_i n_{ij} &= a\sum_i\sum_j x_i n_{ij} + b\sum_i x_i^2 n_{i\cdot}\end{aligned}\right\}$$

Ahora ya estamos en condiciones de resolver el sistema, llamado Sistema de Ecuaciones Normales.

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

4.1.2. Ajuste de una Parábola.

En este caso, la curva seleccionada para ajustarse a la nube de puntos es

$$y_j^* = a + bx_i + cx_i^2$$

y para hallar los parámetros a, b y c debemos minimizar la función

$$f = \sum_i\sum_j \left(y_j - a - bx_i - cx_i^2\right)^2 n_{ij}$$

Las primeras derivadas parciales de la función con respecto a cada uno de los parámetros nos determinan el sistema siguiente:

$$\left.\begin{aligned}\frac{\partial f}{\partial a} &= 2\sum_i\sum_j \left(y_j - a - bx_i - cx_i^2\right)(-1)n_{ij} = 0\\ \frac{\partial f}{\partial b} &= 2\sum_i\sum_j \left(y_j - a - bx_i - cx_i^2\right)(-x_i)n_{ij} = 0\\ \frac{\partial f}{\partial c} &= 2\sum_i\sum_j \left(y_j - a - bx_i - cx_i^2\right)(-x_i^2)n_{ij} = 0\end{aligned}\right\}$$

Realizando las mismas operaciones que en el caso anterior para la recta, el sistema se transforma en

$$\left.\begin{aligned}\sum_j y_j n_{\cdot j} &= aN + b\sum_i x_i n_{i\cdot} + c\sum_i x_i^2 n_{i\cdot}\\ \sum_i\sum_j x_i y_j n_{ij} &= a\sum_i x_i n_{i\cdot} + b\sum_i x_i^2 n_{i\cdot} + c\sum_i x_i^3 n_{i\cdot}\\ \sum_i\sum_j x_i^2 y_j n_{ij} &= a\sum_i x_i^2 n_{i\cdot} + b\sum_i x_i^3 n_{i\cdot} + c\sum_i x_i^4 n_{i\cdot}\end{aligned}\right\}$$

de cuya resolución se obtiene los valores numéricos de los parámetros de la mejor parábola de segundo grado en el sentido mínimo cuadrático para la nube de puntos dada.

4.1.3. Ajuste Hiperbólico.

Son funciones de la forma:

$$yx = b \quad \Leftrightarrow \quad y = b\frac{1}{x}$$

siendo b una constante cualquiera. También se puede considerar la función anterior pero desplazada una cierta cantidad a:

$$y = a + b\frac{1}{x}$$

El ajuste por mínimos cuadrados lo podemos reducir al caso de la recta, ya visto anteriormente, sin más que efectuar la transformación

$$z = \frac{1}{x}$$

con lo que la función se convierte en y = a+bz

4.1.4. Ajuste Potencial.

La forma general de la función potencial es

$$y = a \cdot x^b$$

que de forma análoga al anterior, lo podemos reducir al caso de una recta simplemente tomando logaritmos en ambos miembros.

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

$$y' = a' + bx'$$

4.1.5. Ajuste Exponencial.

La ecuación general es

$$y = a \cdot b^x$$

y repitiendo el caso anterior, al tomar logaritmos se nos reduce al caso de una función lineal.

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

4.2. Método de los Momentos.

El Método de los Momentos se basa en el hecho conocido de que dos distribuciones son tanto más parecidas cuanto mayor cantidad de momentos iguales tengan.

Dada la distribución bidimensional (x_i,y_i) dada por un número N de puntos, recordemos que nuestro objetivo era encontrar una cierta función $y^*=f(x,a_0,...,a_n)$ que se ajuste lo más posible a la nube de puntos obtenida.

Para hallar la función, bastará basarnos en la propiedad que hemos enunciado, para lo cual, sólo tendremos que igualar los n+1 primeros momentos obtenidos de la distribución observada con sus correspondientes de la distribución teórica.

Teniendo en cuenta que los momentos respecto al origen de la distribución observada vienen dados por la expresión:

$$\alpha_{rs} = \frac{1}{N}\sum_{i=1}^{N} x_i^r y_i^s$$

que para el valor s=1 se convierte en:

$$\alpha_{r1} = \frac{1}{N}\sum_{i=1}^{N} x_i^r y_i \quad \forall r:0,1,2,...$$

De forma análoga, los momentos correspondientes a la distribución teórica obtenidos a partir de la función serían:

$$\alpha_{r1} = \frac{1}{N}\sum_{i=1}^{N} x_i^r y_i^* \quad \forall r:0,1,2,...$$

donde, recordemos, $y_i^*=f(x_i,a_0,...,a_n)$

Igualando los n+1 primeros momentos, obtenemos el sistema de ecuaciones:

$$\left.\begin{array}{c} \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i^* \\ \sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} x_i y_i^* \\ \cdots \\ \sum_{i=1}^{N} x_i^n y_i = \sum_{i=1}^{N} x_i^n y_i^* \end{array}\right\}$$

Tenemos así un sistema de n+1 ecuaciones que nos permite calcular el valor de los n+1 parámetros que determinan la función $y_i^*=f(x_i,a_0,...,a_n)$

Si esta función fuera un polinomio de grado k, el sistema que se obtiene es el Sistema de Ecuaciones Normales obtenido por el método de Mínimos Cuadrados cuando la función a ajustar es un polinomio de grado k.

5. REGRESIÓN.

La Regresión tiene por objeto la determinación de una cierta estructura de dependencia que mejor exprese el tipo de relación de la variable Y con las demás. Es decir, trata de poner de manifiesto, a partir de la información disponible, la estructura de dependencia que mejor explique el comportamiento de la variable Y (variable dependiente) a través de todo el conjunto de variables X_1, X_2, ..., X_k (variables independientes) con las que se supone que está relacionada.

En este tema sólo vamos a tratar el caso de disponer de una variable independiente, ya que estamos estudiando distribuciones bidimensionales.

Sean pues, X e Y dos variables cuya distribución conjunta de frecuencias es (x_i, y_j, n_{ij})

DEF Llamaremos Regresión de Y sobre X a la función que explica la variable Y para cada valor de X.

DEF Llamaremos Regresión de X sobre Y a la función que explica la variable X para cada valor de Y.

5.1. Regresión Lineal.

La regresión será lineal cuando la curva de regresión, obtenida o seleccionada, sea una recta. Desarrollaremos este caso particular por ser el más empleado.

5.1.1. Recta de Regresión de Y sobre X.

Haciendo uso de la técnica de mínimos cuadrados para el ajuste de una recta, debíamos hacer mínima la función:

$$f = \sum_i \sum_j \left(y_j - a - bx_i \right)^2 n_{ij}$$

llegando al sistema de ecuaciones normales

$$\left.\begin{aligned} \sum_j y_j n_{\cdot j} &= aN + b\sum_i x_i n_{i\cdot} \\ \sum_i \sum_j y_j x_i n_{ij} &= a\sum_i \sum_j x_i n_{ij} + b\sum_i x_i^2 n_{i\cdot} \end{aligned}\right\}$$

Dividiendo ambas ecuaciones por N, expresamos el sistema en función de los momentos respecto al origen:

$$\left.\begin{aligned} \alpha_{01} &= a + b \cdot \alpha_{10} \\ \alpha_{11} &= a \cdot \alpha_{10} + b \cdot \alpha_{20} \end{aligned}\right\}$$

Para resolver el sistema, multiplicamos la primera ecuación por $-\alpha_{10}$ y sumamos ambas, quedando:

$$b = \frac{\alpha_{11} - \alpha_{10} \cdot \alpha_{01}}{\alpha_{20} - \alpha_{10}^2} = \frac{m_{11}}{m_{20}} = \frac{S_{XY}}{S_X^2}$$

Despejando a en la primera ecuación y sustituyendo el valor de b obtenido

$$a = \alpha_{01} - \frac{S_{XY}}{S_X^2}\alpha_{10} = \bar{y} - \frac{S_{XY}}{S_X^2}\bar{x}$$

Por tanto, la recta de regresión de Y sobre X, en función de los momentos quedará:

$$y = \bar{y} - \frac{S_{XY}}{S_X^2}\bar{x} + \frac{S_{XY}}{S_X^2}x$$

Y reordenando los términos obtenemos:

$$y - \bar{y} = \frac{S_{XY}}{S_X^2}(x - \bar{x})$$

5.1.2. Recta de Regresión de X sobre Y.

Partiendo de la función

$$f = \sum_i \sum_j \left(x_i - a - by_j\right)^2 n_{ij}$$

y realizando un desarrollo análogo al del apartado anterior, llegamos a que la recta de regresión de X sobre Y será:

$$x - \bar{x} = \frac{S_{XY}}{S_Y^2}(y - \bar{y})$$

DEF Llamaremos Centro de Gravedad de la distribución conjunto de X e Y al punto $(\bar{x}, \bar{y})$, lugar donde se cortan ambas rectas de regresión.

5.2. Coeficientes de Regresión.

Los coeficientes de regresión lineal son las pendientes de las rectas de regresión. Así, el coeficiente de regresión de Y sobre X será

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

pero

$$b = \text{tg}\,\alpha = \frac{\Delta y}{\Delta x}$$

luego el coeficiente de regresión de Y sobre X nos mide la tasa de incremento de Y para variaciones de X. Es decir, b indica la variación de la variable Y para un incremento unitario de la variable X.

De forma análoga, el coeficiente de regresión de X sobre Y será

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

y como

$$b' = \text{tg}\,\alpha' = \frac{\Delta y}{\Delta x}$$

b' nos indicará la variación de X correspondiente a un incremento unitario de Y.

Tanto el signo de b como de b' será el signo de la covarianza. Una covarianza positiva nos dará dos coeficientes de regresión positivos y sus correspondientes rectas de regresión crecientes. Si la covarianza es negativa, las dos rectas de regresión serán decrecientes al serlo sus pendientes. En caso de que la covarianza sea cero, las pendientes serán nulas y por tanto las rectas serán paralelas a los ejes coordenados y perpendiculares entre sí.

6. CORRELACIÓN.

DEF Llamamos Correlación al grado de dependencia mutua entre las variables de una distribución.

El problema que se nos plantea es como medir la intensidad con que dos variables de una distribución bidimensional están relacionadas.

Recordemos que a través de la curva de regresión expresábamos la estructura de la relación existente entre las variables, y que para cada valor de x_i obteníamos una diferencia, llamada residuo, entre el valor de Y en la nube de puntos y el correspondiente valor teórico obtenido en la función.

Si todos los puntos de la nube estuvieran en la función, obtendríamos dependencia funcional, siendo el grado de dependencia entre las variables el máximo posible. Cuanto más se alejen los puntos de la función (mayor sean los residuos) menor será la relación entre ambas. Es por ello que para estudiar la dependencia entre las variables vamos a hacer uso de los residuos.

DEF Llamaremos Varianza Residual a la media de todos los residuos elevados al cuadrado.

$$S_{rY}^2 = \sum_i \sum_j \left(y_j - y_j^*\right)^2 \cdot \frac{n_{ij}}{N} = \sum_j \left(y_j - y_j^*\right)^2 \cdot \frac{n_{\cdot j}}{N}$$

Si la varianza residual es grande, entonces los residuos, por término medio, serán grandes, lo que significa que los puntos estarán muy separados de la función, siendo pequeña la dependencia entre las variables. Razonando igual, si la varianza es pequeña, la relación será grande.

Para que la relación entre la dependencia de las variables y la medida utilizada sea directa (en lugar de inversa como ocurre con la varianza residual), definimos el siguiente coeficiente.

DEF Llamamos Coeficiente de Correlación General de Pearson a

$$R = \sqrt{1 - \frac{S_{rY}^2}{S_Y^2}}$$

DEF Llamamos Coeficiente de determinación al cuadrado del coeficiente de correlación, R^2.

6.1. Campo de Variación de R y su Interpretación.

Despejando la varianza residual del coeficiente de correlación R, tenemos

$$S_{rY}^2 = S_Y^2 (1 - R^2)$$

Dado que la varianza marginal de Y y la varianza residual son sumas de sumandos no negativos, ambas serán no negativas, de lo cual deducimos que

$$1 - R^2 \geq 0$$

y de aquí obtenemos

$$-1 \leq R \leq 1$$

Por tanto, el rango de valores de R es el intervalo cerrado [-1,1]

Analicemos ahora, en función de los valores de este coeficiente, que dependencia existe entre las variables.

1) R=1 $\Rightarrow$ $S_{rY}^2 = 0$.

Todos los valores teóricos coinciden con los obtenidos de la observación y, por tanto, la dependencia es funcional. Diremos que existe Correlación Positiva Perfecta, indicando con “Perfecta” que ambas variables varían en el mismo sentido.

2) $R = -1 \Rightarrow S_{rY}^2 = 0$.

En este caso, la dependencia también es funcional pero ambas variables varían en sentidos opuestos. Decimos que existe Correlación Negativa Perfecta.

3) R=0 $\Rightarrow$ $S_{rY}^2 = S_Y^2$

No existe ninguna relación entre la variable Y y la variable X, lo cual significa que no están asociadas. Diremos entonces que la Correlación en Nula.

4) –1<R<0.

La Correlación es Negativa, siendo más fuerte conforme más cerca esté de –1.

5) 0<R<1.

La Correlación es Positiva, y cuanto más próxima esté a uno, más dependencia existirá entre las variables.

6.2. Coeficiente de Correlación Lineal.

Sabemos que la varianza residual es

$$S_{rY}^2 = \sum_i \sum_j \left(y_j - y_j^*\right)^2 \cdot \frac{n_{ij}}{N}$$

y que los valores teóricos de la distribución son

$$y_j^* = \bar{y} + \frac{S_{XY}}{S_X^2}(x_i - \bar{x})$$

Sustituyendo esta expresión en la varianza residual tenemos:

$$S_{rY}^2 = \sum_i \sum_j \left(y_j - y_j^*\right)^2 \cdot \frac{n_{ij}}{N} = \sum_i \sum_j \left(y_j - \left(\bar{y} + \frac{S_{XY}}{S_X^2}(x_i - \bar{x}) \right) \right)^2 \cdot \frac{n_{ij}}{N} =$$

$$= \sum_i \sum_j \left(\left(y_j - \bar{y}\right) - \frac{S_{XY}}{S_X^2}(x_i - \bar{x}) \right)^2 \cdot \frac{n_{ij}}{N} = \sum_i \sum_j \left(y_j - \bar{y}\right)^2 \cdot \frac{n_{ij}}{N} + \frac{S_{XY}}{S_X^2} \sum_i \sum_j \left(x_i - \bar{x}\right)^2 \cdot \frac{n_{ij}}{N} -$$

$$- 2\frac{S_{XY}}{S_X^2} \sum_i \sum_j \left(y_j - \bar{y}\right) \cdot \left(x_i - \bar{x}\right) \cdot \frac{n_{ij}}{N} = S_Y^2 + \frac{S_{XY}^2}{\left(S_X^2\right)^2} S_X^2 - 2\frac{S_{XY}}{S_X^2} \cdot S_{XY} = S_Y^2 - \frac{S_{XY}^2}{S_X^2}$$

DEF Llamamos Coeficiente de Correlación Lineal, al coeficiente de correlación general aplicado al caso de una función lineal. Se denota por r y es

$$r = \sqrt{1 - \frac{S_{rY}^2}{S_Y^2}} = \sqrt{1 - \frac{S_Y^2 - \frac{S_{XY}^2}{S_X^2}}{S_Y^2}} = \sqrt{\frac{S_{XY}^2}{S_X^2 S_Y^2}} = \frac{S_{XY}}{S_X S_Y}$$

Y el Coeficiente de Determinación Lineal es

$$r^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}$$

Ya hemos visto en el apartado anterior que $-1 \le r \le 1$ y la interpretación dada para R también nos sirve para r.

6.3. Interpretación Analítica de r.

Teniendo en cuenta las rectas de regresión de Y sobre X y de X sobre Y y el coeficiente de correlación lineal, podemos expresar las rectas de la siguiente forma:

$$\left.\begin{aligned} y - \bar{y} &= r\frac{S_Y}{S_X}(x - \bar{x}) \\ x - \bar{x} &= r\frac{S_X}{S_Y}(y - \bar{y}) \end{aligned}\right\}$$

Consideremos ahora los siguientes casos:

1) Si r=1.

La varianza residual es cero y los valores teóricos coinciden con los observados. Por tanto, todos los puntos de la nube están en la recta. La correlación lineal es perfecta positiva y las rectas de regresión coinciden, ya que al sustituir r por 1 en la expresión anterior, obtenemos la misma recta. En este caso, la dependencia funcional existente viene reflejada por una recta creciente , al ser la pendiente positiva.

2) Si r = –1.

La correlación es perfecta negativa. En este caso las rectas también coinciden, pero la recta es decreciente al ser la pendiente negativa.

3) Si r=0

La correlación es nula y las dos rectas son

$$\begin{aligned} y - \bar{y} &= 0 \\ x - \bar{x} &= 0 \end{aligned}$$

las cuales son dos rectas paralelas a cada uno de los ejes y por tanto, perpendiculares entre sí.

4) Si –1<r<0.

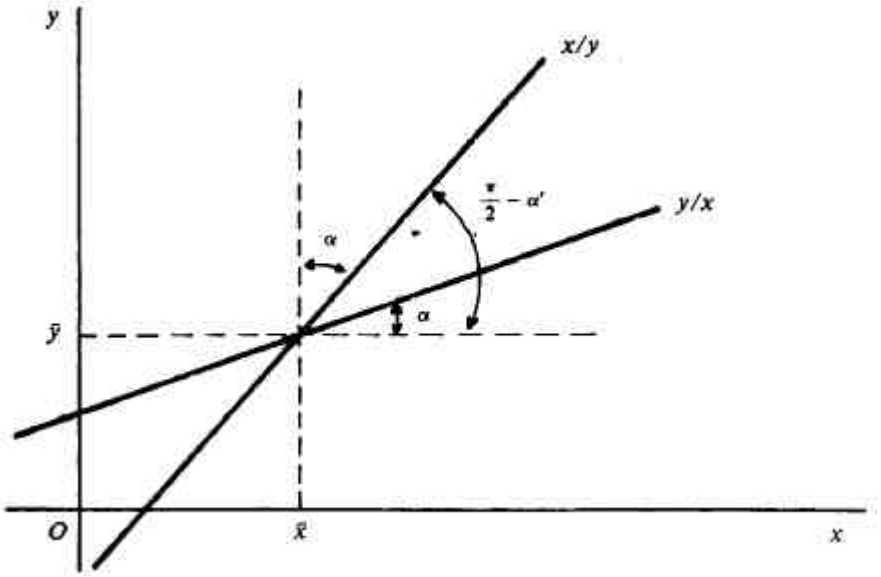
La correlación es negativa y las rectas de regresión, las cuales ahora son diferentes, serán las dos decrecientes ya que el signo de las pendientes es el mismo que el de la covarianza, que es la que da el signo a r.

5) Si 0<r<1

La correlación es positiva, siendo las dos rectas de regresión crecientes.

6.4. Interpretación Geométrica de r.

Partiendo de las rectas de regresión de Y sobre X y de X sobre Y, si las representamos tenemos



Sabemos que

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

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

multiplicando miembro a miembro obtenemos

$$b \cdot b' = \frac{S_{XY}}{S_X^2} \cdot \frac{S_{XY}}{S_Y^2} = \frac{S_{XY}^2}{S_X^2 \cdot S_Y^2} = r^2$$

Es decir

$$r = \sqrt{b \cdot b'}$$

el coeficiente de correlación lineal es la media geométrica de los coeficientes de regresión lineales.

Por otra parte tenemos que

$$b = \operatorname{tg}\alpha$$

$$b' = \operatorname{tg}\alpha' = c\operatorname{tg}\left(\frac{\pi}{2} - \alpha'\right) = \frac{1}{\operatorname{tg}\left(\frac{\pi}{2} - \alpha'\right)}$$

Por lo cual, el coeficiente de correlación lineal se puede expresar como

$$r = \sqrt{b \cdot b'} = \sqrt{\frac{\operatorname{tg}\alpha}{\operatorname{tg}\left(\frac{\pi}{2} - \alpha'\right)}}$$

1) Si r=±1

Tenemos que ambas tangentes son iguales lo cual implica que $\alpha = \frac{\pi}{2} - \alpha'$, lo que geométricamente significa que las dos rectas coinciden, como se puede ver en la gráfica anterior.

2) Si r=0

Entonces tenemos que se debe verificar una de las dos expresiones siguientes:

$$\left.\begin{array}{l} \operatorname{tg}\alpha = 0 \\ \operatorname{tg}\left(\frac{\pi}{2} - \alpha'\right) \to \infty \end{array}\right\} \Rightarrow \begin{array}{l} \alpha = 0 \\ \alpha' \to 0 \end{array}$$

luego las dos rectas son dos paralelas a cada uno de los ejes.

Como conclusión, podemos decir que r es un coeficiente tal que cuando es igual a +1 o –1, el ángulo entre rectas es de cero grados, ya que coinciden. Cuando r es cero, el ángulo formado es de 90º. De aquí deducimos que cuando más se acerque r al valor +1 o –1 mas pequeño será el ángulo entre rectas. Por tanto, r también mide la apertura existente entre las rectas de regresión.

7. VARIANZA DEBIDA A LA REGRESIÓN Y COEFICIENTE DE DETERMINACIÓN LINEAL.

El intento de explicar una variable en función de la otra viene motivado por el supuesto, el cual hemos de comprobar, de que la información que suministra una variable sobre la que se “regresa” va a mejorar el conocimiento del comportamiento de la otra variable. Es decir, se supone que en el caso de la regresión de Y sobre X, Y se explica mejor a través de X que con la distribución marginal de Y.

Para ver en que medida la mejora de la descripción de una variable a través de la otra tiene lugar, vamos a definir primero el concepto de Varianza Debida a la Regresión.

Para ello, hemos de considerar las tres variables que se obtienen en la regresión.

- y_j, que representa los valores observados de la variable Y.
- y_j^* que representa los valores teóricos asignados a x_i en la regresión de Y sobre X.
- e_j que son los residuos o errores que se generan en la regresión mínimo-cuadrática.

Los valores medios de estas tres variables son

1) La media de la serie observada de Y

$$\bar{y} = \sum_i \sum_j y_j \frac{n_{ij}}{N}$$

2) La media de los valores teóricos.

$$\bar{y}^* = \sum_i \sum_j y_j^* \frac{n_{ij}}{N} = \cdots = \bar{y}$$

3) La media de los residuos en la regresión lineal de Y sobre X

$$\bar{e} = \sum_i \sum_j e_j \frac{n_{ij}}{N} = \cdots = 0$$

Teniendo en cuenta estos resultados, podemos definir las siguientes varianzas:

DEF Llamamos Varianza Total de los valores observados a $S_Y^2 = \sum_i \sum_j \left(y_j - \bar{y}\right)^2 \frac{n_{ij}}{N}$

DEF Llamamos Varianza de los Errores o Residuos a

$$S_e^2 = \sum_i \sum_j \left(e_j - \bar{e}\right)^2 \frac{n_{ij}}{N}$$

Si tenemos en cuenta que la media de los residuos tiene valor nulo, al desarrollar la expresión anterior llegamos a que

$$S_e^2 = S_{rY}^2$$

que es la llamada varianza residual.

DEF Llamamos Varianza Debida a la Regresión, a la varianza de los valores teóricos.

$$S_R^2 = \sum_i \sum_j \left(y_j^* - \bar{y}^*\right)^2 \frac{n_{ij}}{N} = \sum_i \sum_j \left(y_j^* - \bar{y}\right)^2 \frac{n_{ij}}{N}$$

En la regresión lineal, podemos encontrar una relación entre las tres varianzas anteriores, la cual pasamos a obtener. Para ello tendremos en cuenta que la regresión es lineal.

$$S_R^2 = \sum_i \sum_j \left(y_j^* - \bar{y}\right)^2 \frac{n_{ij}}{N} = \sum_i \sum_j \left(\bar{y} + \frac{S_{XY}}{S_X^2}(x_i - \bar{x}) - \bar{y}\right)^2 \frac{n_{ij}}{N} =$$

$$= \frac{S_{XY}^2}{\left(S_X^2\right)^2} \sum_i \sum_j \left(x_i - \bar{x}\right)^2 \frac{n_{ij}}{N} = \frac{S_{XY}^2}{\left(S_X^2\right)^2} \cdot S_x^2 = \frac{S_{XY}^2}{S_x^2} = r^2 \cdot S_Y^2$$

Por otra parte, la varianza residual será

$$S_{rY}^2 = S_Y^2(1 - r^2) = S_Y^2 - r^2 S_Y^2$$

Luego tenemos que

$$S_{rY}^2 = S_Y^2 - S_R^2$$
$$S_Y^2 = S_R^2 + S_{rY}^2$$

Dividiendo ambas ecuaciones miembro a miembro

$$1 = \frac{S_R^2}{S_Y^2} + \frac{S_{rY}^2}{S_Y^2}$$

El primer sumando del segundo miembro nos indica la parte de la variación de Y que es explicada por la recta de regresión. El segundo sumando indica la parte no explicada por la recta, la que escapa de ésta o variación residual.

De la expresión anterior tenemos

$$\frac{S_R^2}{S_Y^2} = 1 - \frac{S_{rY}^2}{S_Y^2} = r^2$$

El coeficiente de determinación lineal r^2 nos medirá el grado de acierto de la utilización de la regresión. O lo que es lo mismo, r^2 nos dará el porcentaje de variabilidad de Y que queda explicada por la regresión. La introducción de r, coeficiente de correlación lineal, se justifica con el fin de añadir a r^2 el carácter de la asociación (positiva o negativa).

Teniendo en cuenta la expresión anterior, si $r^2 = 1$, es decir, si la correlación es perfecta:

$$r^2 = \frac{S_R^2}{S_Y^2} = 1 \quad \rightarrow \quad S_R^2 = S_Y^2$$

lo que implica que la varianza residual S_{rY}^2 es nula, luego se ha mejorado al máximo la descripción de Y mediante la utilización de la información suministrada por X. Toda la variabilidad marginal de Y está contenida en la regresión.

Si r=0, caso de correlación nula

$$r^2 = \frac{S_R^2}{S_Y^2} = 0 \quad \Rightarrow \quad S_R^2 = 0 \quad \Rightarrow \quad S_{rY}^2 = S_Y^2$$

es decir, en este caso X no nos sirve para ampliar la descripción del comportamiento de la variable Y.

8. APLICACIONES DE LA REGRESIÓN Y LA CORRELACIÓN.

8.1. Uso y Abuso de la Regresión.

La aplicación de los métodos expuestos de regresión y correlación exige un análisis teórico previo de las posibles relaciones entre las variables. Puede ocurrir que se seleccionen dos variables cualesquiera al azar y que dé la casualidad de que, estadísticamente, la correlación sea perfecta cuando no existe relación posible entre ellas.

Por ejemplo, el hecho de que, casualmente, la correlación lineal entre la tasa de natalidad en Nueva Zelanda y la producción de cereales en España a lo largo de un determinado periodo fuera perfecta no nos debería llevar a suponer que existe algún tipo de relación lineal entre estas variables.

Se deben seleccionar variables entre las que la fundamentación teórica avale algún tipo de relación, evitando, en lo posible, relaciones a través de otra variable principal.. Por ejemplo, el consumo de bebidas puede variar en la misma dirección que el consumo de gasolina, pero no porque una variable dependa directamente de la otra, sino porque ambas van en el mismo sentido que las variaciones de la renta, que será la principal variable explicativa.

8.2. Predicción.

El objetivo último de la regresión es la predicción o pronóstico sobre el comportamiento de una variable para un valor determinado de la otra. Así, dada la recta de regresión de Y sobre X, para un valor $X=x_0$ de la variable, obtenemos y_0

Es claro que la fiabilidad de esta predicción será tanto mayor, en principio, cuanto mejor sea la correlación entre las variables. Por tanto, una medida aproximada de la bondad de la predicción podría venir dada por r.

BIBLIOGRAFÍA RECOMENDADA.

Introducción a la Teoría de la Estadística. Aut.: Mood/Graybill. Ed. Aguilar.

Introducción a la Probabilidad y la Medida. Aut. Procopio Zoroa. Ed. PPU

Algoritmo. Matemáticas II. Cou. Aut.: Vizmanos y Anzola. Edit. SM.

Estadística Teórica y Aplicada. Aut.: A. Vortes. Edit.: PPU

Teoría de Probabilidades y Aplicaciones. Aut.: Cramer. Edit.: Aguilar.