Tema 2: Regresión y Correlación

Descripción Conjunta de Varias Variables

Consideramos el estudio conjunto de dos caracteres de la población, aunque los métodos descritos resultan fácilmente generalizables a un mayor número de variables. Sea:

+ X variable con modalidad $x_1, x_2, \dots$
+ Y variable con modalidad $y_1, y_2, \dots$

$\{(x_i, y_j)\}$

→ La frecuencia absoluta n_{ij} indica el número de veces que se repite el par de valores (x_i, y_j)

→ La frecuencia relativa f_{ij} indica la proporción de veces que se repite la pareja de valores (x_i, y_j) sobre el total de datos de la muestra.

Representaciones

Si el número de observaciones es pequeño, podemos representar las variables en forma de tabla simple.

Var. X	x_1	x_2	...	x_n
Var. Y	y_1	y_2	...	y_n

Representación Tabular Simple

Para poner bastantes observaciones, pero teniendo pocas modalidades.

Var X	Var Y	Frec. Absol.	Frec. Relat.
x_1	y_1	n_1	f_1
x_2	y_2	n_2	f_2
⋮	⋮	⋮	⋮
x_i	y_i	n_i	f_i
⋮	⋮	⋮	⋮
x_k	y_k	n_k	f_k
		N	1

{Distribuciones Marginales}

Son las frecuencias ($n_{i.}$) de los valores de la variable X (sumando por filas) y las frecuencias ($n_{.j}$) de los valores de la variable Y (sumando por columnas)

Representación Tabla Bidimensional

Para poner bastantes observaciones y de modalidades es grande

X \ Y	y_1	y_2	...	y_j	...	y_p	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2p}	$n_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ip}	$n_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kp}	$n_{k.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.p}$	N

Diagrama de Frecuencias (3 variables)

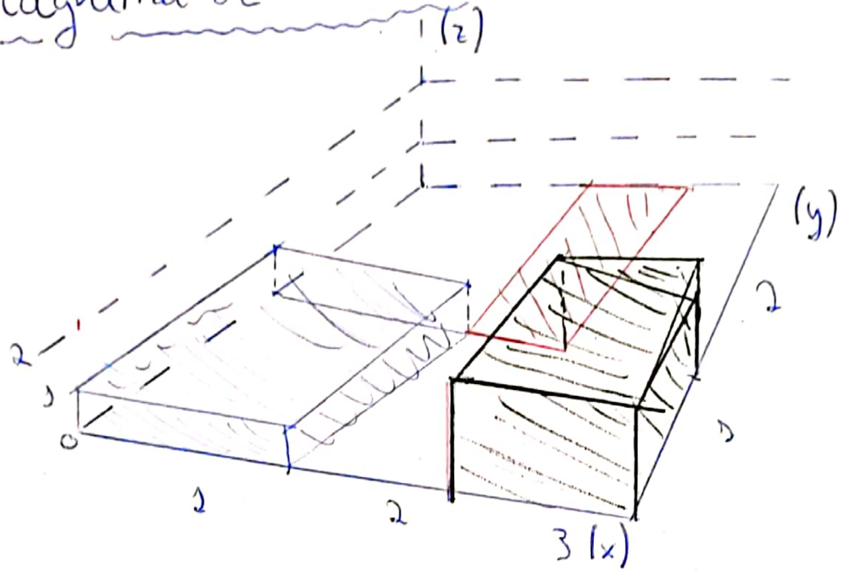
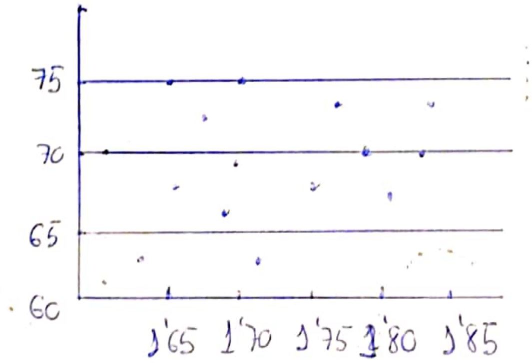
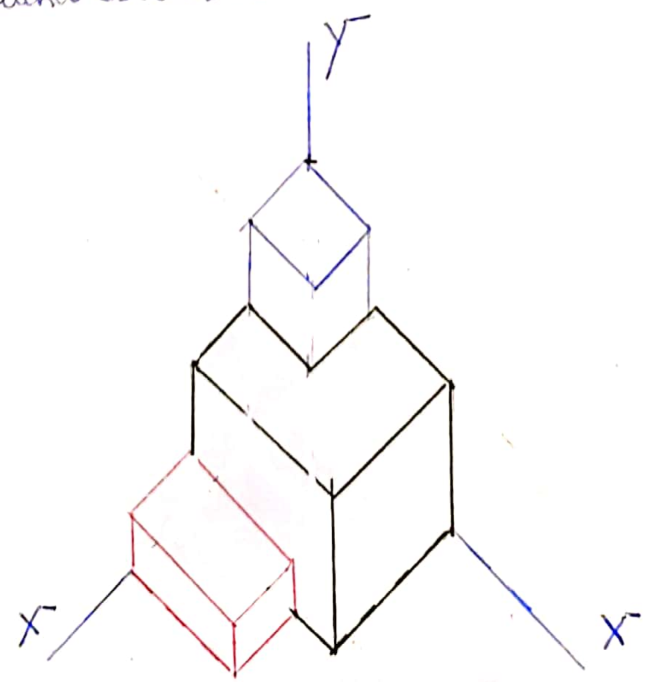


Diagrama de Dispersión



Estereograma

Se usa cuando los datos de ambas variables son o se agrupan en intervalos



$h_{ij} = \frac{n_{ij}}{S_{ij}}$

S_{ij} Área de la modalidad

Frecuencias Marginales

Se obtienen al estudiar una variable con respecto sólo a sí misma, es decir, se realiza el estudio independiente de la otra variable

El nombre de marginal viene dado porque esta frecuencia se obtiene sumando en los márgenes de la tabla de distribución

$$\boxed{n_{i\cdot} = \sum_{j=1}^{k_2} n_{ij}} \xrightarrow{\text{Conllevan a que exista}} \boxed{f_{i\cdot} = \frac{n_{i\cdot}}{N}}$$

$$\boxed{n_{\cdot j} = \sum_{i=1}^{k_1} n_{ij}} \xrightarrow{\text{Conllevan ...}} \boxed{f_{\cdot j} = \frac{n_{\cdot j}}{N}}$$

Distribuciones Condicionadas

Surgen al considerar sólo aquellos valores de la muestra que presentan una determinada modalidad en una de las variables.

Por ejemplo, se llama distribución condicionada del carácter X, respecto a la clase j del carácter Y, y se denota $X/_{y_j}$, a la distribución unidimensional de la variable X, cuando sólo se consideran los individuos de la clase j de Y.

$$\boxed{n_i^j = n_{ij}} \qquad \boxed{f_i^j = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}}$$

Y en el caso contrario, sería cambiando la i por la j.

Momentos

Llamamos momento de orden (r,s) respecto al punto (a,b) a:

$$M_{rs}(a,b) = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i - a)^r (y_j - b)^s \cdot f_{ij}$$

Casos Especiales ⚠

- Momentos Ordinarios (m_{rs}): Cuando $(a,b) = (0,0)$.
- Momentos Centrales (μ_{rs}): Cuando $(a,b) = (m_{10}, m_{01}) = (\bar{x}, \bar{y}) = G$ (Centro de Gravedad)

Llamamos momento ordinario de orden (r,s):

$$m_{rs} = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i)^r (y_j)^s f_{ij}$$

Llamamos momento central de orden (r,s):

$$\mu_{rs} = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i - \bar{x})^r \cdot (y_j - \bar{y})^s f_{ij}$$

Mom. Ordinarios
$$\begin{cases} m_{0,0} = 1 \\ m_{1,0} = \bar{x} = \frac{1}{N} \sum_i n_{i \cdot} x_i \\ m_{0,1} = \bar{y} = \frac{1}{N} \sum_j n_{\cdot j} y_j \end{cases}$$

Mom Centrales
$$\begin{cases} \mu_{0,0} = 1 \\ \mu_{1,0} = 0 \\ \mu_{0,1} = 0 \\ V(x) = m_{2,0} - \bar{x}^2 = \mu_{2,0} = \sigma_x^2 \\ V(y) = m_{0,2} - \bar{y}^2 = \mu_{0,2} = \sigma_y^2 \\ Cov(x,y) = m_{1,1} - \bar{x}\bar{y} = \mu_{1,1} = \sigma_{xy} \end{cases}$$

Relación entre variables

1.- Independencia: No hay relación alguna entre las variables, ninguna proporciona información sobre la otra.

2.- Dependencia funcional: El valor de una variable queda determinado conociendo el valor de la otra variable para esa misma observación.

3.- Dependencia Estadística: Una variable proporciona información sobre la otra, pero conociendo la modalidad de una de ellas no queda determinada la modalidad de la otra.

→ Se dice que el carácter de X es independiente de Y, si:

$$\boxed{f_j^i = f_{i\cdot}} \quad \text{Para todo } i, j$$

→ Se dice que el carácter de Y es independiente de X, si:

$$\boxed{f_i^j = f_{\cdot j}}$$

Aunque tengamos los casos de que las variables pueden ser independientes o tengan dependencia funcional (casos extremos); lo normal que se produce es la dependencia estadística, en la que el conocimiento de una variable da información sobre la otra.

Inciso

Los humanos muchas veces aprendemos en base a este caso de dependencia estadística, es decir, que sin saber la definición de algún objeto lo ligamos a partir de ver muchos casos, por ejemplo, no sabemos lo que es una mesa (definición), pero en base a ver muchas mesas, sabemos que sirve para apoyar objetos.

Machine Learning

En base al inciso anterior comentado, se creó el machine learning que consiste en observar una gran cantidad de individuos o casos en los que se relacionan variables y aprender de ellos.

{Individuos} → Features: características conocidas
{Individuos} → Targets: características que han de predecirse

Algoritmos de Machine Learning

1.- Regresión Lineal

2.- Random Forest (Árboles de Decisión)

3.- Deep Learning (Redes Neuronales)

Deep Learning > Random Forest > Regresión Lineal (Cantidad de Datos para realizarse)

Regresión y Correlación

- Correlación es una medida del grado de dependencia entre las variables.
- Regresión es un método que pretende encontrar un modelo aproximado de la dependencia entre las variables

Representando los datos de la muestra de la variable bidimensional obtenemos una nube de puntos. Se llama línea/curva de regresión a la función que mejor se ajusta a esa nube de puntos.

Si todos los valores de la variable satisfacen la ecuación calculada, se dice que las variables están perfectamente correladas. La ecuación de la curva de regresión nos permite predecir valores desconocidos.

A la vista de la nube de puntos, podemos elegir el tipo de modelo a elegir: lineal, cuadrático, etc.

Ajuste por el método de mínimos cuadrados

Sean los datos $\}x_i, y_i\{$ para dos variables estadísticas X e Y cuantitativas.
El objetivo es encontrar la función $y = f(x)$ de un subconjunto de las funciones reales (rectas, parábolas, hipérbolas, ...) que más se aproxime a los datos.
Se trata pues de minimizar la función objetivo mínimo-cuadrática:

$$\boxed{F = \sum_{i=1}^{k} (y_i - y_i^{est})^2 = \sum_{i=1}^{k} (y_i - f(x_i))^2}$$

⇝ $y_i^{est} = f(x_i)$ es el valor de y estimado por la regresión para x_i.

⇝ $e_i = y_i - y_i^{est}$ es el error cometido por el ajuste para el i-ésimo dato.

Minimizar la función objetivo significa minimizar el Error Cuadrático Medio $\left(ECM = \frac{\sum_i e_i^2}{N}\right)$ y la media cuadrática de los errores $\left(MC = \sqrt{\frac{\sum_i e_i^2}{N}}\right)$

Scanned by CamScanner

Tipos de Ajuste

1.- Ajuste Lineal $y = ax + b$ (2 param)

2.- Ajuste Parabólico $y = ax^2 + bx + c$ (3 param)

3.- Ajuste Hiperbólico $y = \frac{1}{a + bx}$ (2 param)

4.- Ajuste Exponencial $y = ae^{bx}$ (2 param)

Un ajuste de mínimos cuadrados requiere del cálculo de los valores de los parámetros del modelo que minimizan la función objetivo:

$$F(a, b, \dots) = \sum_i (y_i - f(x_i))^2 = \sum_i e_i^2$$

En conclusión, se define la curva general de regresión de Y sobre X como la función que asigna a cada valor x_i de la variable X, la media de la variable $Y/_{x_i}$.

Ajuste Lineal

$$\boxed{y = ax + b} \Longrightarrow F = \sum_{i \in I} (y_i - (ax_i + b))^2$$

Para resolver esta ecuación, necesitamos obtener los parámetros a y b que minimizan la función, esto se resuelve aplicando el gradiente de la función F(vector) e igualándolo a 0.

$$\nabla F = 0 \left\{ \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{array} \right. \Rightarrow \left. \begin{array}{l} -2\sum_i x_i \cdot (y_i - ax_i - b) = 0 \\ -2\sum_i (y_i - ax_i - b) = 0 \end{array} \right\}$$

Ajuste Lineal (Forma Matricial)

Los sistemas de ecuaciones normales, en forma matricial, para el caso de la regresión lineal ~~(xxx)~~ es:

Recta de Y sobre X

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

$$\begin{pmatrix} N & \sum_i y_i \\ \sum_i y_i & \sum_i y_i^2 \end{pmatrix} \cdot \begin{pmatrix} b' \\ a' \end{pmatrix} = \begin{pmatrix} \sum_i x_i \\ \sum_i x_i y_i \end{pmatrix}$$

Ajuste Lineal (Propiedades)

$\boxed{1}$ Y/x

$$\begin{cases} \dfrac{\sum_i y_i}{N} = a \dfrac{\sum_i x_i}{N} + b \\ \dfrac{\sum_i x_i y_i}{N} = a \dfrac{\sum_i x_i^2}{N} + b \dfrac{\sum_i x_i}{N} \end{cases} \Rightarrow \begin{cases} \bar{y} = a\bar{x} + b \\ m_{11} = a m_{20} + b\bar{x} \end{cases}$$

$\boxed{2}$ X/y

$$\begin{cases} \dfrac{\sum_i x_i}{N} = a' \dfrac{\sum_i y_i}{N} + b' \\ \dfrac{\sum_i x_i y_i}{N} = a' \dfrac{\sum_i y_i^2}{N} + b' \dfrac{\sum_i y_i}{N} \end{cases} \Rightarrow \begin{cases} \bar{x} = a'\bar{y} + b \\ m_{11} = a' m_{02} + b'\bar{y} \end{cases}$$

$\rightsquigarrow$ Deducimos que el centro de gravedad $G = (\bar{x}, \bar{y})$ pertenece a ambas rectas. Las rectas X/Y y Y/X se cortan en G. Eliminando a en la de Y/X y a' en la de X/Y:

$$m_{11} - \bar{x}\bar{y} = b(m_{20} - \bar{x}^2) \Rightarrow b = \frac{Cov(x,y)}{\sigma_x^2} = \frac{\mu_{11}}{V(x)}$$

$$m_{11} - \bar{x}\bar{y} = b'(m_{02} - \bar{y}^2) \Rightarrow b' = \frac{Cov(x,y)}{\sigma_y^2} = \frac{\mu_{11}}{V(y)}$$

Coeficiente de Relación de Pearson

Definición

El coeficiente de correlación lineal mide el grado de relación lineal (magnitud y dirección) entre las variables

$$\rho = r = \frac{Cov}{\sigma_x \sigma_y} \qquad (-1 \leq r \leq 1)$$

Casos Particulares

1- $r > 0$ Correlación lineal directa

2- $r < 0$ Correlación lineal inversa

3- $r = 0$ Variables incorreladas

4- $r = 1$ ó $r = -1$ Correlación lineal perfecto (directo o inverso)

Varianza residual. Coeficiente de Determinación

Dada una nube de puntos $\}(x_i, y_i)\{$, llamamos vector residuo $\vec{e} = (e_i)$ con $e_i = y_i - y_{est}$. Es decir, e_i es el error cometido por el ajuste para la i-ésima observación.

Definición

La varianza residual es la varianza del vector residuo.

$$V_r = \sum_i f_i (e_i - \bar{e})^2 = \sum_i f_i \, e_i^2 - \bar{e}^2$$

Definición

El Coeficiente de determinación es:

$$R^2 = 1 - \frac{V_r}{V(y)}$$

El coeficiente de determinación R^2 (caso lineal) verifica $0 \leq R^2 \leq 1$

→ Definición

Llamamos varianza explicada por la regresión a:

$$\boxed{V_e = R^2 V(y)}$$

De $R^2 = 1 - \frac{V_r}{V(y)}$, obtenemos: $V_r = (1 - R^2)V(y)$, luego:

$$V(y) = R^2 V(y) + (1 - R^2)V(y) = R^2 V(y) + V_r = V_e + V_r$$

Así, $R^2 = \frac{V_e}{V(y)}$ representa la fracción de la varianza explicada por el ajuste.

- $R^2 = 1 \Rightarrow$ Ajuste perfecto
- $R^2 = 0 \Rightarrow$ El ajuste no explica nada

Haciendo uso de simplificaciones, podemos llegar a obtener la varianza residual a partir del coeficiente de regresión lineal r:

- $R^2 = r^2$
- $V_r = (1 - r^2)V(y)$

Ajuste Exponencial

$Ln(e^{bx}) = bx \cdot \cancel{Lne}^{1}$

$\boxed{y = ae^{bx}} \Rightarrow Ln(y) = Ln(a) + bx$

Llamando: $Y = Ln(y)$, $A = Ln(a)$, obtenemos: $Y = A + bx$. Podemos ajustar una recta a

$\{(Ln(y_i), x_i)\}$ obteniendo $A = Ln(a)$, $(a = e^A)$ y b que sustituimos en $y = ae^{bx}$

Ajuste Hiperbólico

$$\boxed{y = \frac{1}{b + ax}} \Longrightarrow \frac{1}{y} = ax + b$$

Llamando $Y = \frac{1}{y}$, obtenemos $Y = a + bx$

Podemos ajustar una recta a $\{(\frac{1}{y}, x_i)\}$ obteniendo a, y b que sustituiremos en $y = \frac{1}{ax + b}$

Ajuste Parabólico

$$\boxed{y = ax^2 + bx + c}$$

Podemos deducir las ecuaciones normales minimizando la función:

$$\boxed{F = \sum_i (y_i - (ax^2 + bx + c))^2} \xrightarrow{\text{mediante}} \left| \begin{array}{l} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \\ \frac{\partial F}{\partial c} = 0 \end{array} \right.$$

Pero, esta vez lo vamos a hacer de otra forma:

- Vamos a obtener la función de la forma $f(x) = ax^2 + bx + c \cdot 1$ más próxima a los $\{y_i\}$. Debemos elegir un elemento del subespacio vectorial de las funciones que tienen como base $B = \{x^2, x, 1\}$. Las ecuaciones normales salen de considerar que el vector error debe cumplir:

$$\left\{ \begin{array}{l} (\vec{e}, \vec{x^2}) = 0 \\ (\vec{e}, \vec{x}) = 0 \\ (\vec{e}, \vec{1}) = 0 \end{array} \right. \Longleftrightarrow \left\{ \begin{array}{l} \sum_i (y_i - (ax_i^2 + bx_i + c)) x_i^2 = 0 \\ \sum_i (y_i - (ax_i^2 + bx_i + c)) x_i = 0 \\ \sum_i (y_i - (ax_i^2 + bx_i + c)) = 0 \end{array} \right.$$

$$\Longrightarrow \left\{ \begin{array}{l} \sum_i x_i^2 y_i = a \sum_i x_i^4 + b \sum_i x_i^3 + c \sum_i x_i^2 \\ \sum_i x_i y_i = a \sum_i x_i^3 + b \sum_i x_i^2 + c \sum_i x_i \\ \sum_i y_i = a \sum_i x_i^2 + b \sum_i x_i + cN \end{array} \right.$$

Otros ajustes

Dado un conjunto de puntos $\{(x_i, y_i)\}$

1.-) Ajustar una función del tipo $y = a\,\text{sen}(x) + b\cos(x)$

Una base de las funciones es: $B = \{\text{sen}(x), \cos(x)\}$. Luego se debe cumplir:

$$\begin{cases} (\vec{e}, \overrightarrow{\text{sen}(x)}) = 0 \\ (\vec{e}, \overrightarrow{\cos(x)}) = 0 \end{cases} \Longleftrightarrow \begin{cases} \sum_i (y_i - (a\,\text{sen}(x_i) + b\cos(x_i)))\,\text{sen}(x_i) = 0 \\ \sum_i (y_i - (a\,\text{sen}(x_i) + b\cos(x_i)))\cos(x_i) = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum_i y_i\,\text{sen}(x_i) = a \sum_i \text{sen}^2(x_i) + b \sum_i \cos(x_i)\,\text{sen}(x_i) \\ \sum_i y_i \cos(x_i) = a \sum_i \text{sen}(x_i)\cos(x_i) + b \sum_i \cos^2(x_i) \end{cases}$$

Ajuste de un Plano

$$z = a + bx + cy$$

Dada una nube de puntos $\{(x_i, y_i, z_i)\}_{i \in I}$, podemos deducir las ecuaciones normales minimizando la función.

$$F = \sum_i (z_i - (a + bx_i + cy_i))^2 \quad \text{mediante} \quad \begin{cases} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \\ \frac{\partial F}{\partial c} = 0 \end{cases}$$

Debemos: obtener la función de la forma $z = f(x,y) = a \cdot 1 + b \cdot x + c \cdot y$ y obtener las componentes del elemento (vector $\vec{z}$) del subespacio vectorial de las funciones que tiene como base $B = \{1, \vec{x}, \vec{y}\}$

$$\begin{cases} (\vec{e}, \vec{1}) = 0 \\ (\vec{e}, \vec{x}) = 0 \\ (\vec{e}, \vec{y}) = 0 \end{cases} \Longleftrightarrow \begin{cases} \sum_i (z_i - (a + bx_i + cy_i)) = 0 \\ \sum_i (z_i - (a + bx_i + cy_i))\,x_i = 0 \\ \sum_i (z_i - (a + bx_i + cy_i))\,y_i = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} \sum_i z_i = a \cdot N + b \sum_i x_i + c \sum_i y_i \\ \sum_i z_i x_i = a \sum_i x_i + b \sum_i x_i^2 + c \sum_i y_i x_i \\ \sum_i z_i y_i = a \sum_i y_i + b \sum_i x_i y_i + c \sum_i y_i^2 \end{cases}$$