TEMA 5: REGRESIÓN Y CORRELACIÓN SIMPLE

5.1- INTRODUCCIÓN A LOS MÉTODOS DE AJUSTE.

Partimos de una distribución bidimensional (Xi Yj, n_{ij}) en la que vamos a seguir avanzando estudiando las relaciones entre X e Y→ nos movemos en el campo de la dependencia estadística entre las variables. Para continuar el estudio nos encontramos con el problema del AJUSTE: nos enfrentamos a una "nube de puntos" dada por la representación gráfica en unos ejes de coordenadas de los pares de valores de las 2 variable y buscamos la ecuación que mejor se adapte al conjunto de puntos (obtención de la ecuación de una curva que pase cerca de los puntos dados), imponiéndole determinadas condiciones.

Por tanto en el ajuste habrá dos fases:

1.- Seleccionar el tipo de función que mejor se adapte –gráficamente- al conjunto de datos disponibles, es decir, que mejor represente la relación entre X e Y. La información es la nube de puntos→ útil la representación como primera orientación.

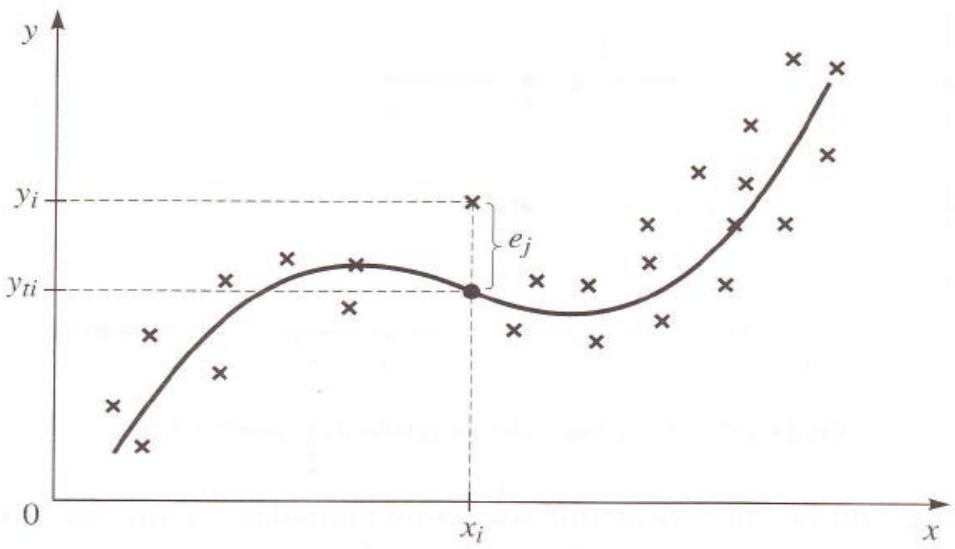
2.- Fijado el tipo de función, a través de su ecuación con un cierto número de parámetros, determinar cuál de las funciones que hay en el plano se adapta mejor al conjunto de puntos (que mejor se ajuste a la nube de puntos de la función).

La determinación de la mejor curva (búsqueda de los parámetros) se consigue imponiendo una serie de condiciones. Según cuáles sean estas condiciones de búsqueda, tendremos uno de los distintos métodos de ajuste existentes.

El principal método de ajuste utilizado (y único que veremos) es el método de NEYMAN o de los MÍNIMOS CUADRADOS.

MÉTODO DE LOS MÍNIMOS CUADRADOS.

De nuestra distribución bidimensional (Xi Yj, n_{ij}) representada en una nube de puntos→ dados los puntos (X_1 Y_1), (X_2 Y_2)…. (X_i Y_j)…. (X_h Y_k), se elige una determinada función de ajuste dada por la expresión siguiente: $Y = f(X, a_1, a_2, \dots, a_n)$ en la que intervienen n parámetros.



Considerando la nube de puntos, al ajustar una función, para cada valor de $X=X_i$ tendremos dos valores de Y:

- El valor observado Y_j correspondiente a la nube de puntos (X_i Y_j)
- El valor teórico Y_{tj} resultado de hacer $X=X_i$ en la función :$Y_{tj} = f(X_i;a_1...a_n) = Y_j^*$

Por tanto, para cada X_i, tendremos la diferencia entre los 2 valores de Y Yj y Y_{tj}, que llamamos RESIDUO = $e_j \rightarrow e_j = Y_j - Y_{tj}$ (diferencia entre Y observado y teórico).

El método de los mínimos cuadrados consiste en la determinación numérica de los parámetros ($a_1...a_n$) de tal manera que los residuos sean mínimos.

$$\min \sum\sum (Y_j - Y_{tj})\, n_{ij} = \min e_j$$

Si tomamos la suma de todos los residuos, se nos presenta el inconveniente de que unos residuos serán de signo positivo y otros de signo negativo, con lo que residuos de distinto signo al sumar se pueden compensar y la suma mínima podría ocultar residuos de cierta importancia a ambos lados de la curva ajustada. Para evitar que los residuos se anulen entre sí, se deberá hacer mínimo la siguiente expresión:

$$\emptyset = \sum_i \sum_j \left(y_j - y_{tj}\right)^2 n_{ij} = \sum_i \sum_j \left(y_j - f(x_i; a_1 a_2 \dots a_n)\right)^2 n_{ij}$$

Al ser los valores teóricos los obtenidos a partir de la función ajustada.

Para hallar de forma única los parámetros $a_1...a_n$ que minimizan $\emptyset$, la condición necesaria es que las primeras derivadas parciales respecto a cada uno de los parámetros se anulen.

$\frac{\partial \emptyset}{\partial a_1} = 2\sum_i \sum_j \left(y_j - f(x_i; a_1 a_2 \dots a_n)\right) \quad n_{ij}(-f'a_1) = 0$

$\frac{\partial \emptyset}{\partial a_2} = 2\sum_i \sum_j \left(y_j - f(x_i; a_1 a_2 \dots a_n)\right) \quad n_{ij}(-f'a_2) = 0$

$\frac{\partial \emptyset}{\partial a_n} = 2\sum_i \sum_j \left(y_j - f(x_i; a_1 a_2 \dots a_n)\right) \quad n_{ij}(-f'a_n) = 0$

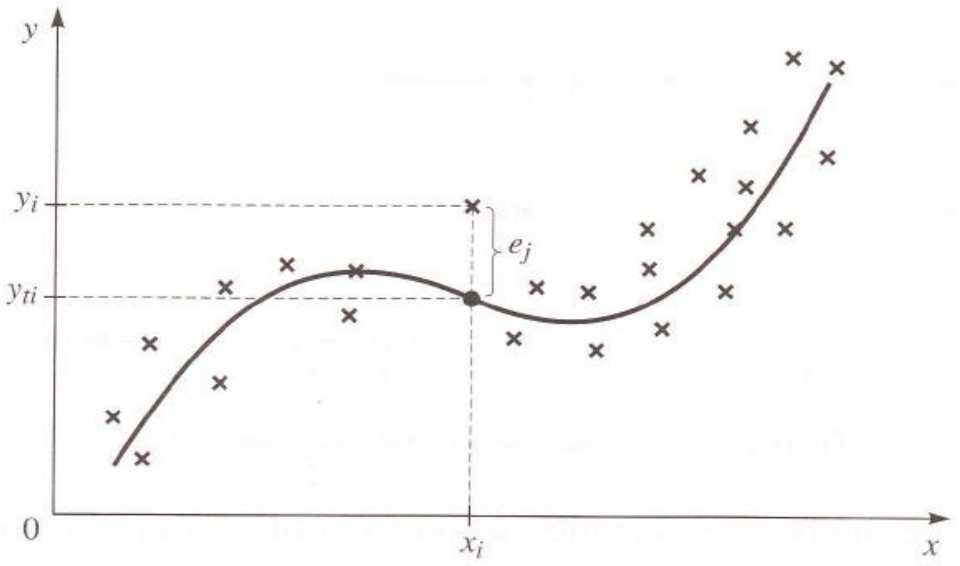
Resolviendo este sistema de ECUACIONES NORMALES queda determinada la función correspondiente y los parámetros.

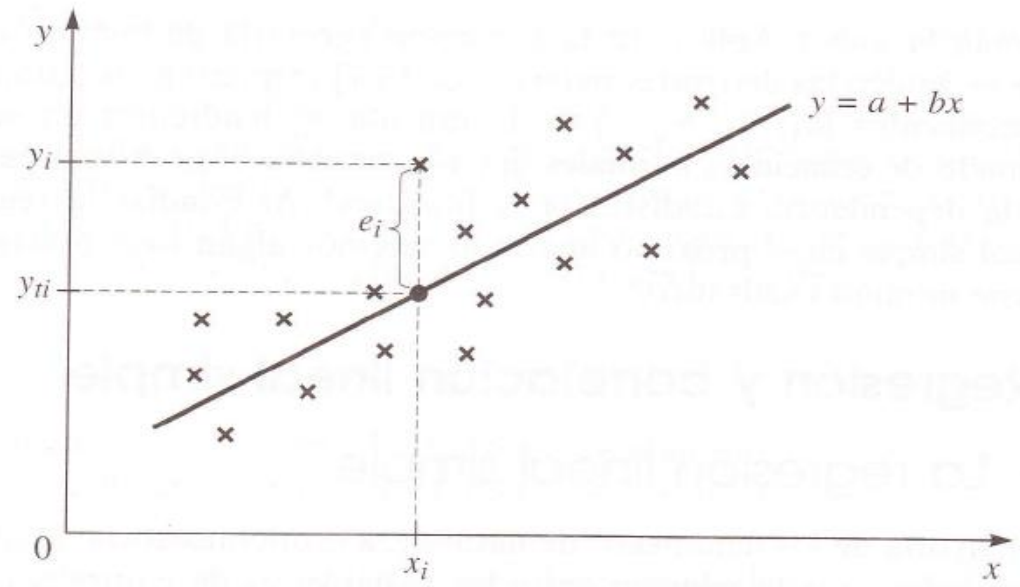
RESUMEN AJUSTE MÍNIMO CUADRÁTICO:

- Se elige una función de ajuste ***y=f(x, a_1 , a_2 ,…, a_n),*** donde intervienen *n* parámetros (***a_1 , a_2 ,…, a_n),***.
- Tenemos que ***y=f(x, a_1 , a_2 ,…, a_n)*** es el valor observado, e ***y*=f(x_i, a_1 , a_2 ,…, a_n)*** es el valor teórico que se obtiene a partir de la curva ajustada.
- Siendo ***(y-y*)=e*** (residuo o error).
- El objetivo es minimizar la suma de los residuos al cuadrado

$$\mathbf{Min\Phi} = \sum_{\mathbf{i=1}}^{r} \sum_{\mathbf{j=1}}^{m} (\mathbf{y_j - y_j^*})^2 \mathbf{n}_{ij}$$

para obtener los *n* parámetros (***a_1 , a_2 ,…, a_n)***.





TIPOS DE AJUSTE

- Ajuste a una recta: ***y=a+bx*** (los resultados son los de regresión lineal).
- Ajuste a una parábola: ***$y=a+bx+cx^2$***
- Ajuste hiperbólico: ***y=a+b(1/x)*** (se utiliza ***z=(1/x)*** y se aplica regresión lineal).
- Ajuste potencial: ***$y=ax^b$*** (tomando log en la ecuación, entonces se aplica un ajuste lineal y se obtienen los resultados aplicando a los parámetros el antilog).
- Ajuste exponencial: ***$y=ab^x$*** (tomando log en la ecuación, entonces se aplica un ajuste lineal y se obtienen los resultados aplicando a los parámetros el antilog).

5.2.- REGRESIÓN.

El fin es encontrar relaciones entre las variables (sucesos a investigar). El investigador intenta traducir esas relaciones en estructuras más manejables, es decir, intenta modelizar esas relaciones funcionalmente a través de un análisis fundamentalmente estadístico (establece relaciones funcionales en donde un número finito de variables X_1,…., X_k se supone que están relacionadas con una variable Y a través de la expresión $Y=f(X_1….X_k)$.

Desde este punto de partida, hay 2 enfoques con que abordar simultáneamente este tema:

1.- Teoría de la CORRELACIÓN (apartado 5.3): estudia el grado de dependencia existente entre las variables.

2.- REGRESIÓN: busca determinar la estructura de dependencia – modelización- que mejor explique el comportamiento de la variable Y (variable DEPENDIENTE o EXPLICADA) en función del conjunto de variables $X_1….X_k$ (variables INDEPENDIENTES O EXPLICATIVAS), con las que se supone está relacionada.

Sean X e Y 2 variables cuya distribución conjunta de frecuencias (Xi Yj, n_{ij}).

Llamamos *Regresión de Y sobre X*: a la función que explica la variable y para cada valor de X , Y=f(X)

Regresión de X sobre Y: comportamiento de X para cada valor de Y X=f(Y)

Para la determinación de las funciones de regresión hay dos criterios diferentes: Regresión I y Regresión II.

REGRESIÓN I:

REGRESIÓN I DE Y SOBRE X:

Considerando la nube de puntos, si nos preguntásemos cual sería el valor de Y para $X=X_i$, existirían varios valores, consideraríamos que sería la media de las Y cuya X sea X_1, es decir, la media de las Y_j cuya abscisa sea X_1 (que no es otra cosa que la media de Y condicionada a que X tome el valor X_1, es decir asigna para cada X_i, un Y_j correspondiente a la media de Y condicionada a $X=X_i$. Los puntos aparecen unidos por una línea para indicarnos que son puntos que pertenecen a una misma regresión.

REGRESIÓN I DE X SOBRE Y: Asigna para cada Y_j, un X_{ti} correspondiente a la media de los X_i condicionados a $Y=Y_j$.

El principal problema de la Regresión I es que está siempre unida por un conjunto de puntos, y no por una curva continua, lo cual lo hace poco deseable para nuestro fin fundamental (explicar una variable a través del comportamiento de la otra). De ahí que se utilice de manera general el criterio de Regresión tipo II.

REGRESIÓN II: por ajuste mínimo-cuadrático:

Base: a través de la información suministrada, cuya representación gráfica es la nube de puntos, 1) se selecciona un tipo de función y posteriormente 2) se ajusta la mejor función de la familia seleccionada aplicando el método mínimo-cuadrático, es decir minimizando los residuos al cuadrado.

Regresión II de Y sobre X:

Se trata de minimizar

$$\sum_i \sum_j \left(y_j - y_{tj}\right)^2 n_{ij} = \sum e_j^2$$

Regresión II de X sobre Y:

Análogamente se trata de minimizar la suma

$$\sum_i \sum_j (x_i - x_{it})^2 n_{ij}$$

Donde X_{it} será el valor teórico de X para un Y_j cualquiera.

La diferencia práctica entre los criterios de Regresión I y Regresión II es que en la Regresión I no fijamos "a priori" el tipo de función, es decir, no seleccionamos ningún tipo de curva, mientras que en la Regresión II esta elección es el primer paso a dar antes de pasar al propio ajuste.

REGRESIÓN LINEAL:

Cuando la curva de regresión obtenida o seleccionada sea una recta.

<u>RECTA DE REGRESIÓN DE Y SOBRE X</u>:

Familia ajustada: recta $Y_{tj} = a + bX_i$ $\quad e_j = Y_j - Y_{tj}$

Aplicar mínimos cuadrados

$$\emptyset = \sum_i \sum_j \left(y_j - y_{tj}\right)^2 n_{ij} = \sum e_j^2$$

Sustituyendo el valor Y_{tj}

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

El sistema de ecuaciones normales sería:

$\frac{\partial \emptyset}{\partial a} = 2 \sum_i \sum_j \left(y_j - a - bx_i\right) \quad (-1) n_{ij} = 0$

$\frac{\partial \emptyset}{\partial b} = 2 \sum_i \sum_j \left(y_j - a - bx_i\right)(-x_i) \quad n_{ij} = 0$

$\sum\sum Y_j n_{ij} = aN + b\sum\sum X_i n_{ij}$

$\sum\sum X_i Y_j n_{ij} = a\sum\sum X_i n_{ij} + b\sum\sum X_i^2 n_{ij}$

Aplicando frecuencias y dividiendo por N, se expresa en función de los momentos respecto al origen.

$$\sum_j y_j \, n_{.j} = aN + b \sum_i x_i n_{i.}$$

$$\sum_i \sum_j x_i \, y_j n_{ij} = a \sum_i x_i n_{i.} + b \sum_i x_i^2 \, n_{i.}$$

$a_{01} = a + b\, a_{10}$

$a_{11} = a\, a_{10} + b\, a_{20}$

Resolvemos el sistema, multiplicando por $(-a_{10})$ la primera ecuación y sumamos ambas

$- \; a_{10}\, a_{01} = - a\, a_{10} - b\, a_{10}^2$

$\quad a_{11} \;\; = \; a\, a_{10} \; + b\, a_{20} \quad +$

$a_{11} - a_{10}a_{01} = b(a_{20} - a_{10}^2)$

$Sxy = b\, S^2{}_x$; $m_{11} = b\, m_{20} \rightarrow b = S_{xy}/S^2{}_x$

Despejando a en la primera ecuación $\quad a_{01} = a + b\, a_{10}$

$a = a_{01} - ba_{10} = \bar{y} - b\bar{x} = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}$

Las estimaciones mínimo cuadráticas de los parámetros a y b son

$b=S_{xy}/S^2_x$ $\quad a = \bar{y}-\frac{S_{xy}}{S_x^2}\bar{x}$

Luego la recta de regresión de Y/X: Y = a + b X

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

Y la recta de regresión de x/y

$X_{ti} = a' + b' Y_j$

$$\emptyset = \sum_i \sum_j (x_i - x_{ti})^2 n_{ij} = \sum_i \sum_j \left(x_i - a' - b' y_j\right)^2 n_{ij}$$

$b' = S_{xy}/S^2_y$ $\quad a' = \bar{x}-\frac{S_{xy}}{S_y^2}\bar{y}$ con lo que la recta sería

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

Estas dos rectas de regresión se cortan en el punto $(\bar{x},\bar{y})$→centro de gravedad de la distribución conjunta.

RESUMEN REGRESIÓN LINEAL:

- El objetivo de la Regresión es encontrar la estructura funcional de dependencia que mejor exprese la relación entre la variable dependiente Y y las variables explicativas o independientes X, Z, etc.

Esquema de Regresión lineal simple

Regresión de Y/X	**Regresión de X/Y**
$y = a + bx$	$x = a' + b' y$
Aplicación del método de Mínimos Cuadrados	
$\min \sum_i \sum_j (y_j - y_j^*)^2 nij$	$\min \sum_i \sum_j (x_j - x_j^*)^2 nij$
Resultados de Mínimos Cuadrados	
$b = \frac{S_{xy}}{S_x^2}$; $a = \bar{y} - b\bar{x}$	$b' = \frac{S_{xy}}{S_y^2}$, $a' = \bar{x} - b'\bar{y}$

Las rectas de Y/X y de X/Y pasan siempre por el punto de las medias, donde se cortan.

En el caso de ser independientes las rectas coinciden con el valor de las medias

COEFICIENTES REGRESION:

Los coeficientes de regresión lineal son las pendientes de las rectas de regresión de Y sobre X.

$b=S_{xy}/S^2_x \qquad b = \operatorname{tg}\alpha = \frac{\Delta y}{\Delta x}$

El coeficiente de regresión de Y/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 X.

Análogamente, el coeficiente de regresión de X sobre Y será $b' = S_{xy}/S^2_y$

$b' = \frac{\Delta x}{\Delta y}$ →variación de x correspondiente a un incremento unitario de Y.

Tanto el signo de b como el de b´será el signo de la covarianza.

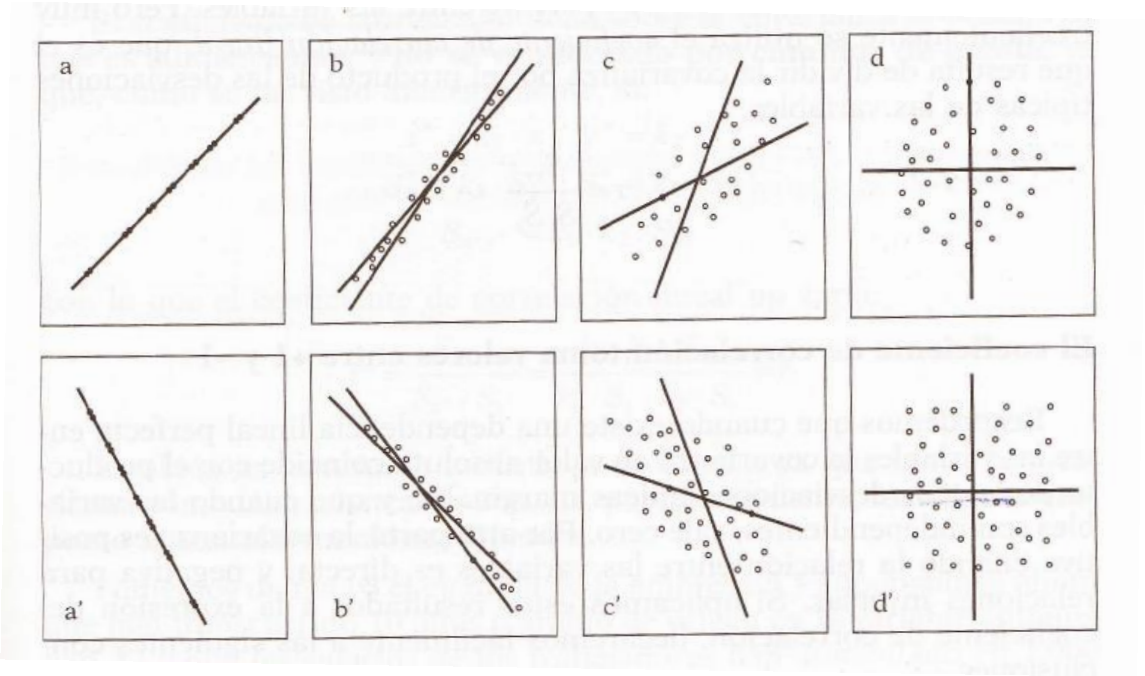
Si Sxy es positiva→b y b´serán positivos y sus correspondientes rectas de regresión positivas.

Si Sxy es negativa→las 2 rectas de regresión serán decrecientes al serlo sus pendientes.

Si Sxy es cero→ b y b´=0, es decir las rectas de regresión serán paralelas a los ejes de coordenadas (y por tanto, perpendiculares entre sí). Resumiendo:

Los coeficientes de regresión a y b tienen la siguiente interpretación:

- El coeficiente "a" es la ordenada en el origen.
- El coeficiente "b" es la tangente, y mide el incremento en Y ante incrementos unitarios en X.



5.3.- CORRELACIÓN

Se llama correlación al grado de dependencia mutua entre las variables. El problema que se plantea será la medición de la intensidad con que dos variables pueden estar relacionadas. Para ello recordemos que a través de la función de ajuste (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, la dependencia sería funcional, y el grado de dependencia sería el máximo posible. Cuanto más se alejen los puntos de la función (mayores serán los residuos) iremos perdiendo intensidad en la asociación. Esto nos indica a utilizar los residuos para medir la dependencia y definimos la *varianza residual* como la media de todos los residuos elevados al cuadrado para evitar que se compensen los residuos.

$$S_{ry}^2 = \sum_i \sum_j (y_j - y_{tj})^2 \frac{n_{ij}}{N}$$

Si S_{ry}^2 es grande→los residuos, por término medio serán grandes→ los puntos estarán alejados de la función y la dependencia será pequeña.

Si S_{ry}^2 es pequeña→la dependencia será grande.

La utilización de la varianza residual presenta el problema de las unidades de medida, lo cual imposibilita la comparación de la dependencia entre grupos de variables. La solución es utilizar el cociente $\frac{S_{ry}^2}{S_y^2}$ siendo S_y^2 la varianza marginal de Y. A mayor dependencia, mayor correlación y así definimos el *coeficiente de correlación general de Pearson*

$$R = \sqrt{1 - \frac{S_{ry}^2}{S_y^2}}$$

El cuadrado del coeficiente de correlación R^2: coeficiente de determinación

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

Campo de variación de R y su interpretación

Despejando la varianza residual $R^2S^2_y = S^2_y - S^2_{ry} \rightarrow S^2_{ry} = S^2_y - R^2S^2_y$

$S^2_{ry} = S^2_y(1-R^2)$

Como S^2_y y S^2_{ry} son sumas de sumandos no negativos, es decir, $S^2_{ry} \geq 0$; $S^2_y \geq 0 \rightarrow 1-R^2 \geq 0$; $R^2 \leq 1 \rightarrow -1 \leq R \leq 1$.

Como $S^2_{ry} = S^2_y(1-R^2)$ veremos lo que ocurre con la dependencia para distintos valores de R:

- Si $R = 1 \rightarrow S^2_{ry}=0$. Todos los valores teóricos coinciden con los observados, los puntos de la nube están en la función, la dependencia es funcional y diremos que existe correlación perfecta positiva, indicando con el calificativo de positivo que ambas variables varían en el mismo sentido.
- Si $R = -1 \rightarrow S^2_{ry} = 0$. La dependencia también es funcional, pero aquí las variables varían en sentidos opuestos y diremos que existe correlación perfecta negativa.
- Si $R = 0 \rightarrow S^2_{ry} = S^2_y$. No conseguimos ninguna explicación de la variable Y relacionándola con la X, las variables X e Y no están asociadas y diremos que la correlación es nula.
- Para $-1<R<0$, la correlación será negativa, siendo más intensa cuanto más próxima esté R a -1.
- Para $0<R<1$, la correlación será positiva y cuanto más próximo esté R a 1 tendremos un mayor grado de asociación.

Coeficiente de correlación lineal:

Nos mide el grado de asociación lineal que existe entre las variables. Correlación está ligado a regresión, es decir hablamos de correlación según una determinada curva de regresión.

Para determinar el coeficiente de correlación lineal partimos del coeficiente de correlación general

$$R = \sqrt{1 - \frac{S^2_{ry}}{S^2_y}}$$

La varianza residual es

$$S_{ry}^2 = \sum_i \sum_j \left(y_j - y_{tj}\right)^2 \frac{n_{ij}}{N}$$

Los valores teóricos en la recta son $Y_{tj} = a + b\, X_i$

Siendo $b=S_{xy}/S^2_x$ $\quad a = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}$

$y_{tj} = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x} + \frac{S_{xy}}{S_x^2}x_i = \bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x})$ Sustituyendo en la varianza residual

$$S_{ry}^2 = \sum_i \sum_j \left(y_j - y_{tj}\right)^2 \frac{n_{ij}}{N} = \sum_I \sum_J \left\{y_j - \left[\bar{y} + \frac{S_{xy}}{S_x^2}\,(x_i - \bar{x})\right]\right\}^2 \frac{n_{ij}}{N}$$

$$= \sum_i \sum_j \left[\left(y_j - \bar{y}\right) - \frac{S_{xy}}{S_x^2}(x_i - \bar{x})\right]^2 \frac{n_{ij}}{N}$$

$$= \sum_i \sum_j \left(y_j - \bar{y}\right)^2 \frac{n_{ij}}{N} + \frac{S_{xy}^2}{(S_x^2)^2} \sum_i \sum_j \left(x_j - \bar{x}\right)^2 \frac{n_{ij}}{N}$$

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

$$= S_y^2 - \frac{S_{xy}^2}{S_x^2}$$

$$S_{ry}^2 = S_y^2 - \frac{S_{xy}^2}{S_x^2}$$

El coeficiente de correlación lineal:

$$r = \sqrt{1 - \frac{S_{ry}^2}{S_y^2}} = \sqrt{1 - \frac{\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right)}{S_y^2}} = \sqrt{\frac{S_y^2 - S_y^2 + \frac{S_{xy}^2}{S_x^2}}{S_y^2}} = \sqrt{\frac{S_{xy}^2}{S_x^2 S_y^2}} = \frac{S_{xy}}{S_x S_y}$$

$$r = \frac{S_{xy}}{S_x S_y}$$

El coeficiente de determinación lineal es :

$$r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

r es un caso particular de R. Su campo de variación será el mismo $\boxed{-1 \leq r \leq 1.}$

Las rectas de regresión de Y sobre X, y de X sobre Y son respectivamente:

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

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

El coeficiente de correlación lineal

$$r = \frac{S_{xy}}{S_x S_y}$$

Podemos relacionar r con los coeficientes de regresión b y b´.

$$b = \frac{S_{xy}}{S_x^2} = \{dividiendo\ y\ multiplicando\ por\ S_y\} = \frac{S_{xy}S_y}{S_y S_x^2} = \frac{S_{xy}S_y}{S_y S_x S_x} = r\frac{S_y}{S_x}$$

$$b' = \frac{S_{xy}}{S_y^2} = r\frac{S_x}{S_y}$$

Con lo que una nueva forma de las rectas de regresión:

$$y-\bar{y} = r\frac{S_y}{S_x}(x-\bar{x})$$

$$x-\bar{x} = r\frac{S_x}{S_y}(y-\bar{y})$$

Casos a considerar:

- Si r = 1, la varianza residual es cero, y los valores teóricos coinciden con los observados, luego todos los puntos de la nube están en la recta, la correlación lineal es perfecta positiva y las rectas de regresión coinciden, al sustituir r por 1 quedarían:

$$y-\bar{y} = \frac{S_y}{S_x}(x-\bar{x})$$
$$x-\bar{x} = \frac{S_x}{S_y}(y-\bar{y})$$

que es la misma recta. En este caso la dependencia funcional existente viene reflejada por una recta creciente, ya que la pendiente es positiva.

Cuando la correlación lineal es perfecta las dos rectas de regresión coinciden (rectas crecientes y con pendientes inversas):

$$r = \frac{S_{xy}}{S_x S_y}$$

$$b = \frac{S_{xy}}{S_x^2}$$

$$b' = \frac{S_{xy}}{S_y^2}$$

$$bb' = \frac{S_{xy}}{S_x^2}\frac{S_{xy}}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2$$

$$r = \frac{S_{xy}}{S_x S_y} = \sqrt{bb'}$$

Si r = 1 $\quad 1 = \sqrt{bb'} \quad$ 1=bb´ b=1/b´

- Si r = -1, la correlación es perfecta negativa. Aquí, también coinciden las rectas, pero las pendientes son negativas y las rectas son decrecientes.

$$y - \bar{y} = -\frac{S_y}{S_x}(x - \bar{x})$$

$$x - \bar{x} = -\frac{S_x}{S_y}(y - \bar{y})$$

- Cuando r = 0, la correlación es nula y las dos rectas son:

$y - \bar{y} = 0 \rightarrow y = \bar{y}$

$x - \bar{x} = 0 \rightarrow x = \bar{x}$

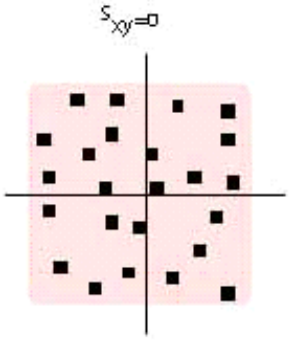
2 rectas paralelas, cada una de ellas respecto a un eje, es decir, perpendiculares entre sí con punto de corte $(\bar{x}\ \bar{y})$. Si tomamos la recta de regresión de y/x que es $y = \bar{y}$, por mucho que varíe X, la variable Y no varía, con lo que el grado de asociación es nulo. Igualmente se reproduce en la recta de regresión x/y.

- Para -1<r<0 la correlación será negativa, y las rectas de regresión que ahora serán distintas y las 2 decrecientes. Si r es negativo, la covarianza será negativo y tanto b como b´ serán negativos.
- Para 0<r<1 la correlación es positiva, y las 2 rectas de regresión crecientes.

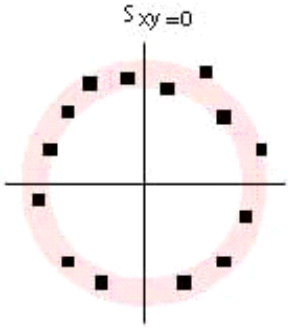
Correlación lineal e independencia estadística

Si dos variables eran independientes estadísticamente su covarianza era cero, ahora se añade que la correlación lineal es cero, es decir variables incorrelacionadas linealmente. No se da la viceversa, ya que la correlación lineal entre las variables puede ser nula, al serlo la covarianza y no ser independientes ya que la covarianza se puede anular sin que se cumpla la condición de independencia. Las variables pueden estar incorrelacionadas linealmente y ser dependientes, ya que al ser r=0, lo único que asegura es que la asociación lineal es nula, pero esas variables pueden estar relacionadas según otro tipo de asociación (parabólica, exponencial..)

El coeficiente de correlación lineal es invariante ante cambios de origen y de escala.



Las dos variables son independientes.



Hay dependencia entre las dos variables, aunque la covarianza sea nula.

Propiedades de r

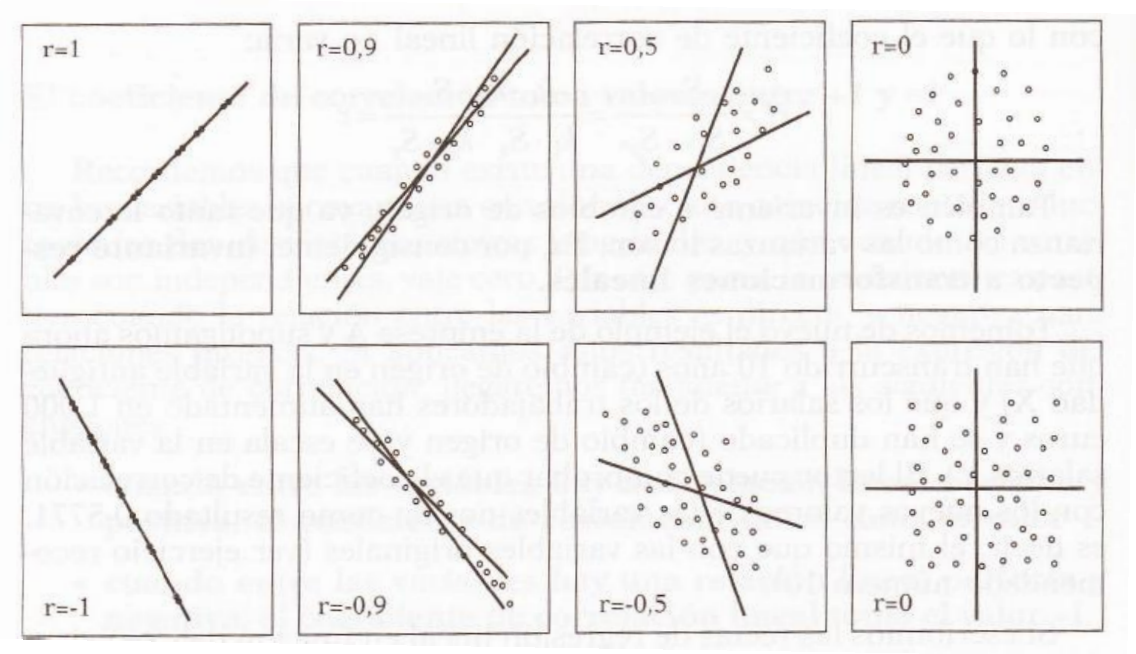
- Es adimensional
- Sólo toma valores en [-1,1]
- Las variables son independientes ⇔ r=0
- Relación lineal perfecta entre dos variables ⇔ r=+1 o r=-1
- Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.
- tiene el mismo signo que S_{xy} por tanto de su signo obtenemos el que la posible relación sea directa o inversa.
- r es útil para determinar si hay relación lineal entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)

Resumiendo, **el coeficiente de correlación lineal**:

Mide el grado de dependencia lineal entre las variables

Coeficiente de determinación y correlación lineal (mide el grado de relación lineal entre las variables)
$$r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = bb',\ 0 \le r^2 \le 1$$ $$r = \frac{S_{xy}}{S_x S_y} = \sqrt{bb'},\ -1 \le r \le 1$$ Si r=0 las series están incorreladas, pero no implica independencia estadística.

El coeficiente r de correlación lineal no se ve afectado por cambios de origen o escala.



5.4.- VARIANZA DEBIDA A LA REGRESIÓN Y COEFICIENTE DE DETERMINACIÓN LINEAL.

Y = a + b X, donde X es la variable independiente e Y la variable dependiente. ¿En qué grado X explica Y?.

El intento de explicar una variable en función de la otra viene motivado por el supuesto de que la información que suministra la 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/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 previamente el concepto de VARIANZA DEBIDA A LA REGRESIÓN: S^2_R . Para ello, consideramos las 3 variables que se obtienen en la regresión:

- Y_j: representa a la serie de valores observados de Y.
- Y_{tj}: conjunto de valores teóricos asignados a cada X_i en la regresión de Y/X.
- e_j: conjunto de residuos o errores que se generan en la regresión mínimo-cuadrática.

Los valores medios de estas 3 variables son:

- La media de la serie observada de Y

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

- La media de los residuos en la regresión lineal de Y/X

$$\bar{e} = \sum_i \sum_j e_j \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - y_{tj}) \frac{n_{ij}}{N}$$

$$\bar{e} = 0$$

En la regresión lineal, la suma de los residuos es nula, y su media, por lo tanto, también.

- La media de los valores teóricos

$$\bar{y_t} = \sum_i \sum_j y_{tj} \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - e_j) \frac{n_{ij}}{N} = \bar{y} - \bar{e} = \bar{y}$$

El valor medio de los valores teóricos coincide con el valor medio de los valores observados.

Podemos definir las siguientes varianzas:

- La varianza total de los valores observados.

$$S_y^2 = \sum_i \sum_j (y_j - \bar{y}) \frac{n_{ij}}{N}$$

- La varianza de los errores o residuos

$S_e^2 = \sum_i \sum_j (e_j - \bar{e}) \frac{n_{ij}}{N} = \sum_i \sum_j e_j^2 \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - y_{tj})^2 \frac{n_{ij}}{N} = S_{ry}^2$ que es la llamada varianza residual.

- Por último, la VARIANZA DEBIDA A LA REGRESIÓN, que es la varianza de los valores teóricos.

$$S_R^2 = \sum_i \sum_j (y_{tj} - \bar{y_t})^2 \frac{n_{ij}}{N} = \sum_i \sum_j (y_{tj} - \bar{y})^2 \frac{n_{ij}}{N}$$

Es la media de las desviaciones al cuadrado de los valores teóricos respecto a la media de la distribución observada. Entre estas 3 varianzas existe en la regresión lineal una relación.

Vamos a calcular S^2_R teniendo en cuenta que: los valores teóricos en la recta son $Y_{tj} = a + b X_i$

Siendo $b=S_{xy}/S^2_x$ $\quad a = \bar{y}-\frac{S_{xy}}{S_x^2}\bar{x}$

$y_{tj} = \bar{y}-\frac{S_{xy}}{S_x^2}\bar{x}+ \frac{S_{xy}}{S_x^2}x_i = \bar{y}+ \frac{S_{xy}}{S_x^2}(x_i - \bar{x})$ con lo que sustituyendo en la varianza debida a la regresión:

$$S_R^2 = \sum_i \sum_j \left(y_{tj} - \bar{y}\right)^2 \frac{n_{ij}}{N} = \sum_i \sum_j \left[\bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x}) - \bar{y}\right]^2 \frac{n_{ij}}{N}$$

$$= \frac{S_{xy}^2}{(S_x^2)^2} \sum_i \sum_j (x_i - \bar{x})^2 \frac{n_{ij}}{N} = \frac{S_{xy}^2}{(S_x^2)^2} S_x^2 = \frac{S_{xy}^2}{S_x^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} S_y^2 = r^2 S_y^2$$

Por otra parte, la varianza residual era:

$S^2_{ry} = S^2_y(1-r^2) = S^2_y - r^2\, S^2_y$

$S^2_{ry} = S^2_y - S^2_R$ $\qquad$ $\boxed{S^2_y = S^2_R + S^2_{ry}}$

Es decir, la varianza marginal, que nos mide la variación de y en la distribución marginal observada, se puede de descomponer en la suma de 2 varianzas:

- S^2_R : la varianza debida a la regresión, la dispersión de los valores en la recta.
- S^2_{ry} : la varianza residual, que mide las desviaciones entre los valores observados y los teóricos.

Dividiendo ambos miembros por S^2_y

$\boxed{S^2_y = S^2_R + S^2_{ry}}$ $\qquad$ $1 = \frac{S_R^2}{S_y^2} + \frac{S_{ry}^2}{S_y^2}$

$\frac{S_R^2}{S_y^2}$: *parte de la variación de Y que es explicada por la recta de regresión.*

$\frac{S_{ry}^2}{S_y^2}$: la parte que no es explicada por la recta, la que escapa de ésta o variación residual.

$\frac{S_R^2}{S_y^2} = 1 - \frac{S_{ry}^2}{S_y^2} = r^2$

El coeficiente de determinación lineal r^2 nos medirá el grado de acierto de la utilización de la regresión. Nos da el porcentaje de variabilidad de Y que queda explicada por la regresión.

- Si $r^2 = 1$ (si la correlación es perfecta)

$$r^2 = \frac{S_R^2}{S_y^2} = 1 \; y \; por \; tanto \; S_R^2 = S_y^2$$

La varianza residual $S^2_{ry} = 0$. Se ha mejorado al máximo la descripción de Y mediante la utilización de la información suministrada por X.

- Si r = 0, caso de correlación nula.

$$r^2 = \frac{S_R^2}{S_y^2} = 0 \; y \; por \; tanto \; S_R^2 = 0 \; y \; S_{ry}^2 = S_y^2$$

La variable x no nos sirve para ampliar la descripción del comportamiento de la variable Y.

5.5.- APLICACIONES DE LA REGRESIÓN Y LA CORRELACIÓN

- USO Y ABUSO DE LA REGRESIÓN (peligro relaciones espúreas entre las variables). La aplicación de los métodos de regresión y correlación exige un estudio exhaustivo de las posibles relaciones entre las variables. Puede ocurrir que se relacionen 2 variables cualesquiera que no tengan nada que ver y que de la casualidad de que desde el punto de vista estadístico exista correlación perfecta pero desde el punto de vista teórico no se pueden relacionar ni realizar ningún estudio coherente. Por ejemplo si estudiamos la producción de patata en España con el número de neumáticos fabricados en una empresa estadounidense en un determinado periodo de tiempo y resulta que obtenemos una correlación perfecta entre estas variables , este resultado no nos debe llevar a suponer que existe algún tipo de relación entre ellas. Lo que se debe hacer es elegir variables que desde el punto de vista teórico exista algún tipo de relación, como por ejemplo la demanda de consumo depende de nuestra renta disponible.
- 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 si la recta de regresión de Y/X es:

$y = \bar{y} + \frac{S_{xy}}{S_x^2}(x - \bar{x})$ o bien Y = a + b X

la predicción de y para $X=X_0$ será:

$y_0 = \bar{y} + \frac{S_{xy}}{S_x^2}(x_o - \bar{x})$ o bien $Y_0 = a + b X_0$

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 aproximativa de la bondad de la predicción viene dada por r^2 (coeficiente de determinación lineal).

Resumiendo:

Descomposición de la varianza marginal de la variable dependiente	
Regresión de Y/X	**Regresión de X/Y**
$S_y^2 = S_{ry}^2 + S_{Ry}^2$ S_{ry}^2 (varianza residual, debida a causas ajenas a la regresión) S_{Ry}^2 (varianza debida a la regresión, varianza explicada por la variable independiente)	$S_x^2 = S_{rx}^2 + S_{Rx}^2$ S_{rx}^2 (varianza residual, debida a causas ajenas a la regresión) S_{Rx}^2 (varianza debida a la regresión, varianza explicada por la variable independiente)

$$S_{ry}^2 = \sum_i \sum_j (e_j - \bar{e})^2 \frac{n_{ij}}{N} = \sum_i \sum_j (e_j)^2 \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - y_j^*)^2 \frac{n_{ij}}{N} = S_y^2 - \frac{S_{xy}^2}{S_x^2} = S_y^2 - \frac{S_{xy}^2}{S_x^2}\frac{S_y^2}{S_y^2} = S_y^2\left(1-r^2\right)$$

$$S_{Ry}^2 = \sum_i \sum_j (y_j^* - \bar{y}^*)^2 \frac{n_{ij}}{N} = \sum_i \sum_j (y_j^* - \bar{y})^2 \frac{n_{ij}}{N} = \frac{S_{xy}^2}{S_x^2}\frac{S_y^2}{S_y^2} = S_y^2 r^2$$

$$S_{rx}^2 = \sum_i \sum_j (e_i - \bar{e})^2 \frac{n_{ij}}{N} = \sum_i \sum_j (e_i)^2 \frac{n_{ij}}{N} = \sum_i \sum_j (x_i - x_i^*)^2 \frac{n_{ij}}{N} = S_x^2 - \frac{S_{xy}^2}{S_y^2}\frac{S_x^2}{S_x^2} = S_x^2\left(1-r^2\right)$$

$$S_{Rx}^2 = \sum_i \sum_j (x_i^* - \bar{x}^*)^2 \frac{n_{ij}}{N} = \sum_i \sum_j (x_i^* - \bar{x})^2 \frac{n_{ij}}{N} = \frac{S_{xy}^2}{S_y^2}\frac{S_x^2}{S_x^2} = S_x^2 r^2$$

Demostraciones en regresión lineal:

Media de los residuos:

$$\bar{e} = \sum_i \sum_j e_j \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - y_j^*)\frac{n_{ij}}{N}$$

$$sustituyendo \quad y_j^* = \bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x})$$

$$\bar{e} = \sum_i \sum_j (y_j - [\bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x})])\frac{n_{ij}}{N} =$$

$$= \sum_i \sum_j (y_j - \bar{y})\frac{n_{ij}}{N} - \frac{S_{xy}}{S_x^2}\sum_i \sum_j (x_i - \bar{x})\frac{n_{ij}}{N} = 0 + \frac{S_{xy}}{S_x^2}0 = 0$$

Media de los valores teóricos:

$$\bar{y}^* = \sum_i \sum_j y_j^* \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - e_j) \frac{n_{ij}}{N} = \bar{y} - \bar{e} = \bar{y}$$

Varianza residual:

$$S_{ry}^2 = \sum_i \sum_j (y_j - y_j^*)^2 \frac{n_{ij}}{N}, \quad sustituyendo \quad y_j^* = \bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x})$$

$$S_{ry}^2 = \sum_i \sum_j \left[(y_j - \bar{y}) - \frac{S_{xy}}{S_x^2}(x_i - \bar{x}) \right]^2 \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - \bar{y})^2 + \frac{S_{xy}^2}{S_x^4} \sum_i \sum_j (x_i - \bar{x})^2 -$$

$$-2 \frac{S_{xy}}{S_x^2} \sum_i \sum_j (y_j - \bar{y})(x_i - \bar{x}) = S_y^2 + \frac{S_{xy}^2}{S_x^4} S_x^2 - 2 \frac{S_{xy}}{S_x^2} S_{xy} = S_y^2 - \frac{S_{xy}^2}{S_x^2}$$

Varianza debida a la regresión:

$$S_{Ry}^2 = \sum_i \sum_j (y_j^* - \bar{y})^2 \frac{n_{ij}}{N}, \quad sustituyendo \quad y_j^* = \bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x})$$

$$S_{Ry}^2 = \sum_i \sum_j (\bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x}) - \bar{y})^2 \frac{n_{ij}}{N} = \frac{S_{xy}^2}{S_x^4} \sum_i \sum_j (x_i - \bar{x})^2 \frac{n_{ij}}{N} = \frac{S_{xy}^2}{S_x^4} S_x^2 = \frac{S_{xy}^2}{S_x^2}$$

Bondad del ajuste (grado de asociación o de dependencia

Se mide mediante el coeficiente de determinación general

Regresión de Y/X:

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2}, 0 \le R^2 \le 1$$

Regresión de X/Y:

$$R^2 = 1 - \frac{S_{rx}^2}{S_x^2}, 0 \le R^2 \le 1$$

$$S_{ry}^2 = \sum_i \sum_j (e_j - \overline{e}_j)^2 \frac{n_{ij}}{N} = \sum_i \sum_j ((y_j - y_j^*) - \overline{e}_j)^2 \frac{n_{ij}}{N} = S_y^2 (1 - R^2)$$

$$\overline{e}_j = \sum_i \sum_j e_j \frac{n_{ij}}{N} = \sum_i \sum_j (y_j - y_j^*) \frac{n_{ij}}{N}$$

$$S_{rx}^2 = \sum_i \sum_j (e_i - \overline{e})^2 \frac{n_{ij}}{N} = \sum_i \sum_j ((x_i - x_i^*) - \overline{e}_i)^2 \frac{n_{ij}}{N} = S_x^2 (1 - R^2)$$

$$\overline{e}_i = \sum_i \sum_j e_i \frac{n_{ij}}{N} = \sum_i \sum_j (x_i - x_i^*) \frac{n_{ij}}{N}$$

Regresión de Y/X:

$$R^2 = 1 - \frac{S_{ry}^2}{S_y^2} = \frac{S_{Ry}^2}{S_y^2}, 0 \le R^2 \le 1$$

Regresión de X/Y:

$$R^2 = 1 - \frac{S_{rx}^2}{S_x^2} = \frac{S_{Rx}^2}{S_x^2}, 0 \le R^2 \le 1$$

En el caso de **regresión lineal simple se cumple:**

$$R = r$$

Pero si tenemos **regresión no lineal** o más de una variables explicativa (**regresión lineal múltiple**), la relación ya no se cumple:

$$R \neq r$$

Regresión lineal múltiple

$$y_j^* = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi}$$

Donde los valores observados son: $y_j = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi} + e_j$

Ajuste por mínimos-cuadrados:

$$\mathbf{Min\Phi} = \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{y_j - y_j^*})^2 \mathbf{n}_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{y_j - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi})^2 n}_{ij}$$

Para ello derivamos parcialmente la expresión anterior respecto a cada uno de los p+1 coeficientes (b0,b1, .., bp) e igualamos a cero. Obtenemos un sistema de p+1 ecuaciones normales, para determinar p+1 incógnitas (los p+1 coeficientes), que habrá que resolver:

Interpretación de los coeficientes:

- El coeficiente bo es el valor que tomaría la variable dependiente cuando todas las independientes valen cero.

- El coeficiente b1 es el incremento que se produce en la variable dependiente ante incrementos unitarios en la variable X1, manteniendo constante el resto de variables.

- Análogamente para el resto de coeficientes.

R≠r