Regresión lineal simple

Regresión lineal simple

1.- Introducción

El coeficiente de correlación (r_{xy}), estudiado en los puntos anteriores, permitía conocer la magnitud de la relación (supuestamente lineal) existente entre dos variables. En el presente apartado nos introduciremos en el concepto de regresión lineal, donde estudiaremos la estructura de relación existente entre tales variables. Ambos conceptos -regresión y correlación- están íntimamente ligados, mientras el primero especifica la forma de la relación, el segundo, sobre la base de esta forma, estudia la intensidad de la relación establecida.

De una manera más concreta, mediante el modelo de regresión especificaremos la *ecuación de regresión* que nos permitirá un doble objetivo: a) describir de una manera clara y concisa la relación existente entre ambas variable y b), predecir los valores de una variable en función de la otra.

En un sentido muy amplio, y hablando en términos puramente estadísticos, podemos afirmar que el análisis de regresión es un método que permite analizar la variabilidad de una determinada variable en función de la información que le proporcionan una o más variables (Pedhazur, 1982). Se concreta, como hemos indicado, en el estudio de relación entre variables, de forma tal que una determinada variable -variable respuesta, explicada, dependiente o criterio- pueda expresarse en función de otra u otras variables - predictoras, explicativas, independientes o regresores-, lo que permitirá predecir los valores de la variable respuesta en función de las variables explicativas, así como determinar la importancia de éstas. Por otro lado, se especifica que la estructura de la relación es lineal. Este aspecto es importante por cuanto se descartan aquí otros tipos de relaciones. Por esta razón, con cierta frecuencia nos referiremos a la regresión lineal como *modelo de regresión lineal*, en el sentido de que se aplica una cierta concepción -modelo- que tenemos de la realidad merced a la cual se supone que las relaciones entre variables sigue una cierta estructura -la estructura lineal.-.

Hemos de decir, aunque sólo sea por curiosidad histórica, que el término "regresión" se debe a Sir Francis Galton (1822-1911) estudiando la relación de la estatura entre padres e hijos. Observó que los padres altos tenían hijos altos, aunque no tan altos como sus progenitores. Igualmente, los padres bajos tendían a tener descendencia de baja estatura aunque más altos que sus respectivos padres. En ambos casos, pues, existía una cierta tendencia a la estatura media, o dicho en términos de propio Galton, existía una "regresión a la mediocridad". Aunque hoy día el término de "regresión lineal" está muy lejos de sus primeras intenciones ha quedado así acuñado, aunque con otros propósitos.

Es evidente el interés el modelo de regresión lineal aplicado a Ciencias Humanas y de la Salud, donde no podemos encontrar relaciones exactas como ocurre en otras áreas de la ciencia, pero sí ciertas tendencias susceptibles de ser cuantificadas. Supóngase, por citar tan

sólo algunos posibles casos de estudio, el efecto de una cierta terapia sobre las respuestas de los pacientes sometidos a ella, los gastos de publicidad de una empresa y el consumo ciudadano, el efecto del tabaco sobre el cáncer, el clima laboral y la productividad en una empresa o la calidad de enseñanza y el rendimiento académico. En todos ellos hay algún aspecto de la conducta que nos interesa prever (y en última instancia, controlar). Merced a la ligazón que presenta la conducta con alguna variable relevante (y que se entiende manipulable por el investigador) podemos ejercer algún tipo de control sobre aquella interviniendo sobre la variable que incide sobre la misma. De esta forma, lograremos nuestros propósitos en cuanto a salud, por ejemplo, eliminado el consumo de tabaco, o bien una determinada terapia cognitivo-conductual se mostrará efectiva en la remisión de la depresión.

Como se ha indicado, en el presente capítulo, nos limitaremos al estudio de la regresión donde se estudia la relación que sobre la variable de respuesta ejerce una única variable explicativa. Este tipo de regresión -la más sencilla de las posibles- se denomina por esta razón *regresión lineal simple*.

2.- Diagrama de dispersión

Previo a todo análisis, resulta conveniente una primera inspección visual de los datos al objeto de comprobar la conveniencia o no de utilizar el modelo de regresión simple. Se recurre a este respecto, a la representación conjunta de los datos mediante el *diagrama de dispersión* o *nube de puntos*. Una simple ojeada nos permitirá determinar (se entiende *grosso modo*): a) si existe relación o no entre las variables y b) si ésta es o no lineal. Además pueden extraerse otras informaciones de interés, como son: c) el grado de estrechez de la nube de puntos, indicadora de la intensidad de la relación, d) si existen valores anómalos que distorsionan la posible relación, o e), si la dispersión de los datos a lo largo de la nube de puntos es uniforme, lo que tendrá su importancia, tal como veremos en los próximos apartados.

La información obtenida es importante para encarar la actuación más conveniente. Una nube redondeada y sin contornos definidos (fig. *2a*) es indicadora de ausencia de relación. La variable explicativa es irrelevante y no merece la pena seguir con el modelo en cuestión. En la figuras *2b* y *2c* se sugiere una relación lineal, más fuerte en la figura *2b*, debido a su mayor estrechez, aunque en ambos casos un análisis estadístico posterior se hace necesario para confirmar con seguridad la relación insinuada en los gráficos. En la figura *2d* la relación es claramente curvilínea (como ocurre si relacionamos ansiedad con rendimiento) y no procede a aplicar el modelo lineal de regresión. Aquí podemos optar por transformar los datos a efecto de lograr linealidad, o lo que puede ser más conveniente, respetar los datos y elaborar el modelo pertinente. En la figura *2e*, la dispersión no es constante a lo largo del recorrido de los datos -*heterocedasticidad*-, lo que imposibilita, como se tendrá ocasión de comprobar, la aplicación del modelo lineal de regresión. Por otro lado, en la figura *2f* un par de datos anómalos -*outliers*- ejercen una distorsión importante sobre el modelo, lo que obligará a replantearse la conveniencia de eliminarlos o bien incluirlos en el modelo, con la consiguiente transformación del mismo.

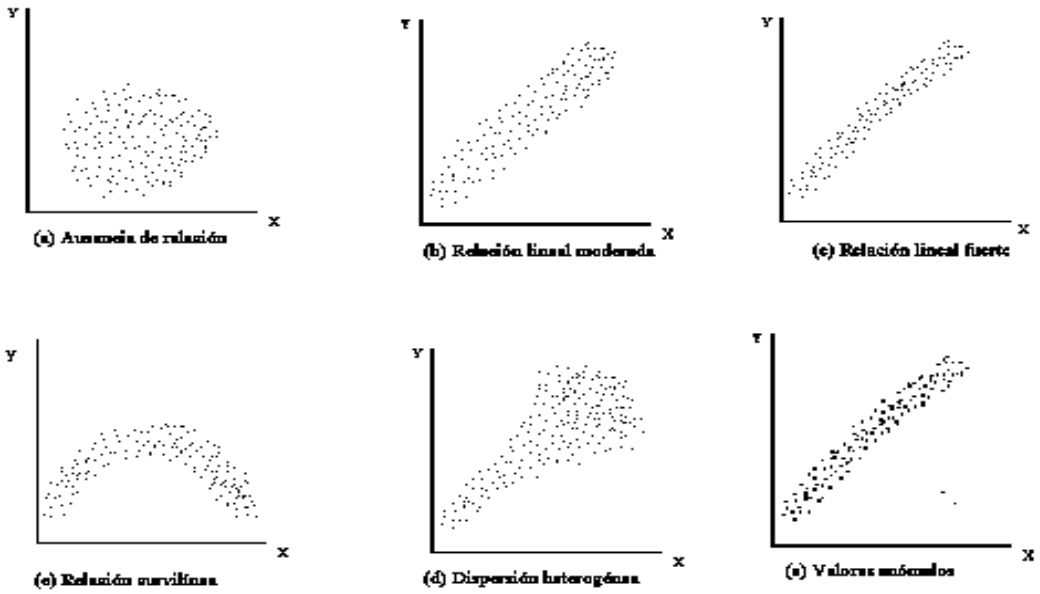


Figura 2. Posibles diagramas de dispersión

3.- Especificación del modelo de regresión lineal simple

Como se ha observado anteriormente, cuando existen razones para suponer la existencia de una relación lineal entre dos variables, podremos establecer la siguiente estructura de relación:

$$Y = \alpha + \beta X + \varepsilon$$

En términos gráficos, esta relación quedaría expresada mediante el siguiente diagrama causal:

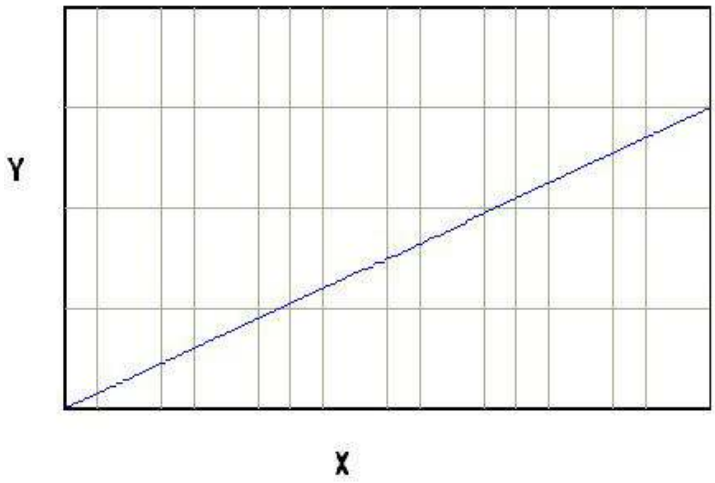
donde podemos distinguir las variables *X*, *Y* y ε . La variable *X*, origen de la flecha en la figura 1.1 es la variable observada cuya incidencia sobre Y deseamos estudiar. En términos

de la ecuación (1.1) es la variable que sirve de base para la predicción. Se le denomina variable *predeterminada*, *explicativa*, *predictora*, *independiente*, *exógena* o simplemente, *regresor*. En nuestra opinión, variable explicativa o predictora, son los términos cuyos significados ilustran mejor el propósito de estas variables. Se dice que es *fija* si sus valores son establecidos por el investigador; por ejemplo, cuando analizamos el efecto que el número de miligramos de una determinada droga tiene sobre el tiempo de reacción a ciertos estímulos visuales y fijamos previamente los valores de X . Por el contrario, se dice que es aleatoria cuando sus valores no están determinados por el investigador sino que se elige una muestra aleatoria de sujetos y se miden ambas variables. Por ejemplo, si queremos investigar la relación entre inteligencia y rendimiento en matemáticas en niños de 14 años, y para ello, seleccionamos una muestra de la población de niños de 14 años, midiendo, posteriormente, su nivel de inteligencia y su rendimiento en matemáticas. Los valores de inteligencia obtenidos son el resultado de la medición en la muestra (modelo de efectos aleatorios para X), pero la muestra estudiada no viene condicionada por valores predefinidos de inteligencia.

La variable Y, punto final de la flecha, es la variable que el modelo pretende dar cuenta Se la suele denominar como variable de *respuesta*, *explicada*, *dependiente*, *criterio* o *endógena*. El objeto de la regresión va a ser, precisamente, estimar la relación que Y presenta con X y predecir sus valores en sujetos no medidos en la muestra. Igualmente, en nuestra opinión, consideramos más conveniente el término de variable de *respuesta* o *explicada.*

La variable ε representa el componente de error en la predicción de la variable Y debido la relación estocástica entre Y y X. Se le denomina entre otros nombres como *error*, *perturbación*, o *residual*. Debe su valor fundamentalmente a dos tipos de factores: a) medición incorrecta de la variable Y, y b) influencia de otras variables omitidas por el modelo. Si salimos del esquema determinista que impera en Ciencias Humanas y concedemos un cierto valor al azar y a la espontaneidad habremos de añadir a los puntos anteriores un tercer punto: c) variabilidad inherente a la conducta humana.

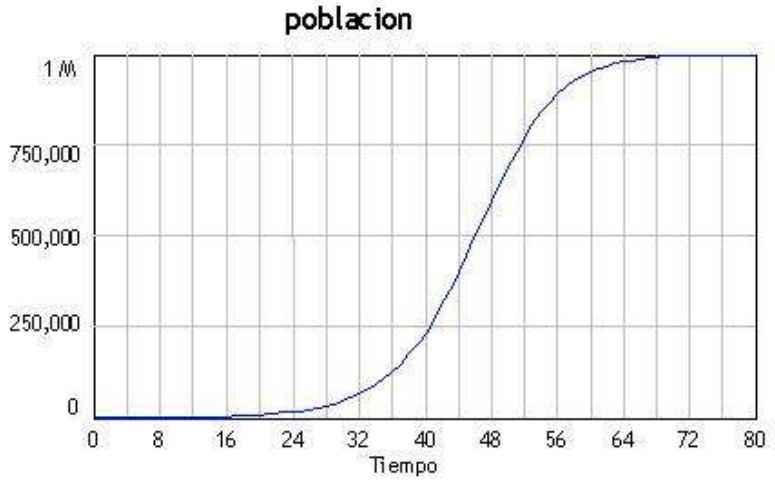
Es importante destacar que aquí nos ocupamos de relaciones entre variables exclusivamente lineal; esto es, de variables cuya estructura de relación es del tipo:



No obstante pueden establecerse otras muchas formas de relación posible que no son abordables directamente desde el planteamiento lineal. Por ejemplo, es bien conocida que la relación entre ansiedad y rendimiento sigue una U invertida:

Un pequeño incremento en los valores de ansiedad sirve para activar al sujeto, y en consecuencia, aumenta su rendimiento, mientras que a partir de un cierto punto, la ansiedad desarbola a dicho individuo impidiéndole concentrarse adecuadamente.

Otro tipo de relación no lineal puede observarse en la evolución de un determinado rumor a lo largo del tiempo, que sigue una relación sigmoidal:



En los momentos iniciales hay poca gente con conocimiento del tema, lo que hace que la extensión del rumor sea pequeña. Conforme aumenta el número de sujetos conocedores de tal rumor hay más posibilidad de interacción con las personas desconocedoras del tema, con lo que hay una gran progresión, hasta llegar a un cierto punto en el que casi toda la población está saturada y son ya pocos los individuos que restan por enterarse de la

cuestión, de forma tal que el incremento es cada vez más reducido, hasta alcanzar el valor de cero, cuando el rumor ha llegado a extenderse por toda la población. Este tipo de fenómenos es muy conocido en biología, especialmente en dinámica de poblaciones, característico de la evolución de una cierta población con recursos limitados.

Otro ejemplo. La relación entre esfuerzo y aprendizaje no es lineal sino tal como se expone en la siguiente figura:

En el comienzo, pequeños esfuerzos supone un aprendizaje relativamente rápido, pero a partir de un determinado punto grandes esfuerzos no se ven proporcionalmente compensados (recuérdese a este respecto, el aprendizaje de idiomas).

No daremos más ejemplos para no aburrir al lector. Tan sólo dejar constancia de que la relación lineal es una de las relaciones posibles y de que existen otros modelos alternativos. El modelo lineal es bien conocido y existe una extensa literatura al respecto, por lo que es frecuentemente utilizado, aunque no siempre con las debidas precauciones. Ya veremos más adelante, en el tema correspondiente al análisis de residuos cómo tratar toda la casuística de modelos supuestamente no lineales.

3.1.- Supuestos del modelo

El modelo de regresión lineal simple para la población establece como hipótesis estructural básica lo siguiente:

$$Y = \alpha + \beta X + \varepsilon$$

la puntuación de un sujeto en la variable criterio Y depende linealmente de la puntuación del sujeto en la variable predictora X más una perturbación o error ε. Otra forma de expresar el modelo es:

$$Y = \hat{Y} + \varepsilon$$

donde la puntuación Y predicha por el modelo de regresión es:

$$\hat{Y} = \alpha + \beta X$$

De la expresión (1.15) se deduce que el error en la predicción será:

$$\varepsilon = Y - \hat{Y}$$

Los parámetros de la ecuación (1.14) -ecuación de regresión verdadera- (α y β) son generalmente desconocidos y han de ser estimados a partir de los valores observados en una muestra de sujetos. Para que las inferencias a la población -estimación- así como los contrastes de hipótesis acerca de los parámetros sean adecuados es necesario que las variables implicadas cumplan las siguientes características estadísticas:

(a) *Linealidad*. El primer supuesto establece que el *valor esperado* (media) en la variable *Y* para cada uno de los valores *X* se encuentra sobre la recta de regresión "verdadera" de *Y* sobre *X*, o dicho de otra manera, la recta de regresión de *Y* sobre *X* vendrá determinada por los valores medios de *Y* para cada valor de *X*. En consecuencia, la esperanza matemática de los errores será cero. Así:

$$E(Y \mid X) = \alpha + \beta X$$

En términos de los errores:

$$E(\varepsilon) = 0$$

Ya que:

$$E(\varepsilon) = E(Y - \hat{Y}) = E(Y - \overline{Y}) = E(Y) - E(\overline{Y}) = \overline{Y} - \overline{Y} = 0$$

(b) *Homocedasticidad*. El segundo supuesto establece que las varianzas de *Y* para cada valor de *X* son todas iguales σ^2, esto es, la dispersión de la variable *Y* a todo lo largo de la recta de regresión es constante. El interés de esta propiedad reside en la ventaja de utilizar un único valor para todo el recorrido de *X* a la hora de estimar valores de *Y* a partir de *X*, lo que otorga simplicidad al modelo. Así pues:

$$Var(Y \mid X_i) = \sigma^2$$

Obsérvese que la distribución de los errores es la misma que la de la variable dependiente en torno a la recta de regresión (para valores fijos de X). En consecuencia, su varianza coincidirá con la de los errores ya que en la expresión $Y = \alpha + \beta X + \varepsilon$ la variabilidad en *Y* para un cierto valor de *X* lo aporta ε:

$$Var(Y \mid X_i) = E(Y_i - \hat{Y}_i)^2 = E(Y_i - \alpha - \beta X_i)^2 = E(\varepsilon^2) = \sigma_\varepsilon^2$$

c) *Ausencia de autocorrelación*. El tercer supuesto establece que las variables aleatorias *Y* son independientes entre sí; es decir, la covarianza (o bien, correlación) entre dos valores de *Y* cualesquiera es cero. Cuando los valores de *Y* hacen referencia a sujetos

distintos -estudios transversales- esta propiedad suele cumplirse. Otro caso sucede en estudios longitudinales donde se efectúan diferentes mediciones de los mismos sujetos a lo largo del tiempo, y que por razones de inercia suelen presentar autocorrelación. Así:

$$Cov(Y_i Y_j) = 0$$

O bien:

$$Cov(\varepsilon_i \varepsilon_j) = 0$$

d) *Normalidad de las distribuciones*. Este supuesto establece que la forma de la distribución de *Y* para cada valor de *X* sigue una ley normal. Se cumple, entonces, la condición de normalidad. Esta propiedad, junto a la condición de homocedasticidad facilita la inferencia estadística del valor de Y poblacional a partir del valor de X. Así:

$$Y_i \approx N(\mu_{y|x}, \sigma^2_{y|x})$$

Y en término de los errores:

$$\varepsilon_i \approx N(0, \sigma^2)$$

Hay que decir en relación a este supuesto que le modelo de regresión es bastante robusto frente a violaciones del mismo. Por otro lado, para tamaños de muestras grandes, el *teorema central del límite* garantiza su cumplimiento.

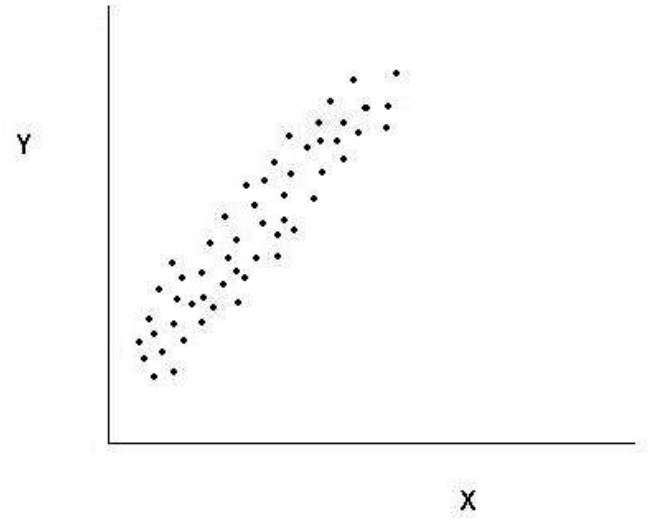
Además de estos requisitos necesarios a efectos de inferencia estadística y contrastes de hipótesis han de respetarse otros supuestos relacionados con el modelo de regresión en cuanto modelo descriptivo. Estos son:

(a) *El modelo ha de estar correctamente especificado*, lo que implica el doble cometido de no haber excluido variables independientes relevantes y el no haber incluido variables independientes irrelevantes. Este requisito cumple su verdadera dimensión en la regresión múltiple donde las variables independientes han de ser seleccionadas cuidadosamente. Cuando se trata de una única variable independiente, la precaución ha de cifrarse en esa variable y aquí la evidencia es palpable si el modelo no ha sido correctamente especificado.

(b) *La variable independiente ha de haber sido medida sin error*. Se quiere decir con ello que las puntuaciones empíricas obtenidas en *X* son precisamente sus puntuaciones verdaderas. Este requisito es un tanto ideal ya que el error de medida está implícito en toda medición. A este respecto hay que decir que en modelos más completos (*Modelos Estructurales*) se contempla la fiabilidad en la medida. Obsérvese por otro lado, que la exactitud en la medición no es requisito para la variable *Y*, ya que esta circunstancia queda contemplada en el error γ.

4.- Estimación de parámetros

Los datos observados en una determinada muestra presentan una configuración del tipo:



denominado *diagrama de dispersión* o bien *nube de puntos*. Dicha configuración carece de operatividad matemática. No obstante, según el modelo convenido, la estructura de relación entre X e Y se supone lineal. Así pues, hemos de determinar la recta:

$$\hat{Y} = \alpha + \beta X$$

que mejor represente la nube de puntos correspondiente a la muestra observada, y cuyos valores (a y b) sean buenos estimadores de la verdadera ecuación de regresión (α y β):

$$E(Y \mid X) = \alpha + \beta X$$

referente a la población de origen.

Podríamos utilizar varios métodos en la determinación de la recta que mejor ajuste a la mencionada nube de puntos. Todos ellos tendrán, obviamente, como objetivo fundamental reducir al mínimo el error global cometido, lo que se traduce, de alguna forma, en minimizar el conjunto de errores e obtenido para el total de las observaciones. A este respecto, podríamos establecer el siguiente criterio:

$$\sum_{i=1}^{N} e_i \Rightarrow mínimo$$

Este procedimiento presenta el inconveniente de que puede lograrse una suma de cero existiendo grandes errores positivos y negativos que quedarían neutralizados entre sí. Esta situación podríamos solventarla con dos procedimientos: a) operando con los valores

absolutos de los errores:

$$\sum_{i=1}^{N} | e | \Rightarrow mínimo$$

O bien, b) elevando al cuadrado tales valores:

$$\sum_{i=1}^{N} e_i^2 \Rightarrow mínimo$$

De estos dos procedimientos, el último, denominado criterio de mínimos cuadrados es el preferible. Varias razones lo avalan:

a) El hecho de elevar al cuadrado las puntuaciones no solamente resuelve el problema del signo, sino que además magnifica los errores grandes, lo cual obliga aún más a reducir tales errores.

b) Algebraicamente entraña menos dificultades operar con sumas de cuadrados que con sumas de valores absolutos.

c) Y por último, y este es el punto más importante, las estimaciones de los parámetros de la ecuación de regresión (a y b) obtenidas mediante el criterio de los mínimos cuadrados son estimaciones sin sesgo, y por el teorema de Gauss-Markov presentan la mínima varianza (ver al respecto el Apéndice A). Además, las estimaciones obtenidas mediante mínimos cuadrados son coincidentes con las logradas por el procedimiento de máxima verosimilitud.

4.1.- Estimación mediante mínimos cuadrados

a) Puntuaciones directas.

En lo que sigue demostraremos, dado un conjunto de datos ofrecidos en puntuaciones directas, que la ecuación de la recta $\hat{Y} = a + bX$ cuyo ajuste sigue el criterio de los mínimos cuadrados es aquella que tiene por pendiente:

$$b = r_{xy} \frac{S_y}{S_x}$$

y de ordenada en el origen:

$$a = \overline{Y} - b\overline{X}$$

Efectivamente, tengamos la expresión:

$$\sum_{i=1}^{N} e_i^2 \Rightarrow mínimo$$

Sustituyendo los errores por su valor:

$$\sum_{i=1}^{N} e^2 = \sum_{i=1}^{N} \left(Y - \hat{Y}\right)^2 = \sum_{i=1}^{N} (Y - (a + bX))^2 = \sum_{i=1}^{N} \left(Y^2 + (a + bX^2) - 2Y(a + bX)\right)$$

Esta función tendrá un mínimo para los valores que anulen la primera derivada respecto a *a* y *b*. Así pues, calculemos primeramente la derivada parcial respecto a *a*. Haciendo operaciones tenemos:

$$\frac{\delta\left(\sum_{i=1}^{N} e^2\right)}{\delta a} = a + b\overline{X} - \overline{Y} = 0$$

De donde:

$$a = \overline{Y} - b\overline{X}$$

Para calcular *b* procedamos de igual manera. Igualemos a cera la derivada parcial respecto a *b*, y haciendo operaciones:

$$\frac{\delta\left(\sum_{i=1}^{N} e^2\right)}{\delta b} = b\left(\frac{\sum_{i=1}^{N} X^2}{N} - \overline{X}^2\right) - \frac{\sum_{i=1}^{N} XY}{N} - \overline{Y}\overline{X} = 0$$

Despejando b:

$$b = \frac{\frac{\sum_{i=1}^{N} XY}{N} - \overline{X}\overline{Y}}{\frac{\sum_{i=1}^{N} X^2}{N} - \overline{X}^2} = \frac{S_{xy}}{S_x^2} = \frac{r_{xy} S_x S_y}{S_x^2} = r_{xy} \frac{S_y}{S_x}$$

b) puntuaciones centradas

Tengamos la ecuación de regresión en directas:

$$\hat{Y} = a + bX$$

Sustituyamos *a* por su valor:

$$\hat{Y} = a + bX = (\overline{Y} - b\overline{X}) + bX = \overline{Y} - b\overline{X} + b\overline{X} = \overline{Y} + b(X - \overline{X})$$

Donde se nos indica que el valor pronosticado en *Y* es precisamente su media (el valor previsto en ausencia total de información) más el efecto de la variable *X*.

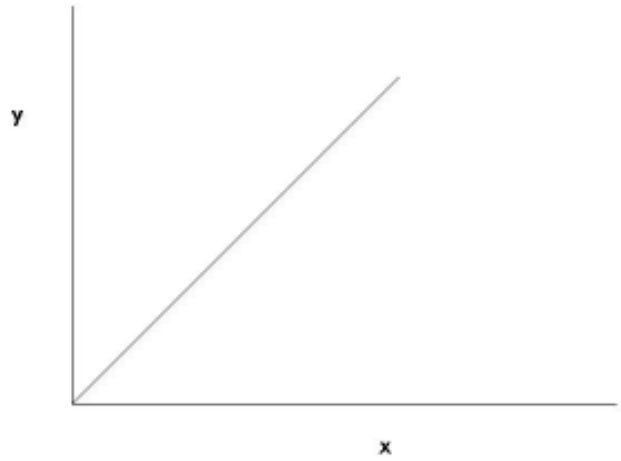
Ahora, si pasamos $\overline{Y}$ al primer miembro de la ecuación:

$$\hat{Y} - \overline{Y} = b(X - \overline{X})$$

Se observa en el primer miembro las puntuaciones centradas de Y y en el segundo las puntuaciones centradas de X. Sustituyendo, entonces, por la notación adecuada el modelo en puntuaciones centradas queda:

$$\hat{y} = bx$$

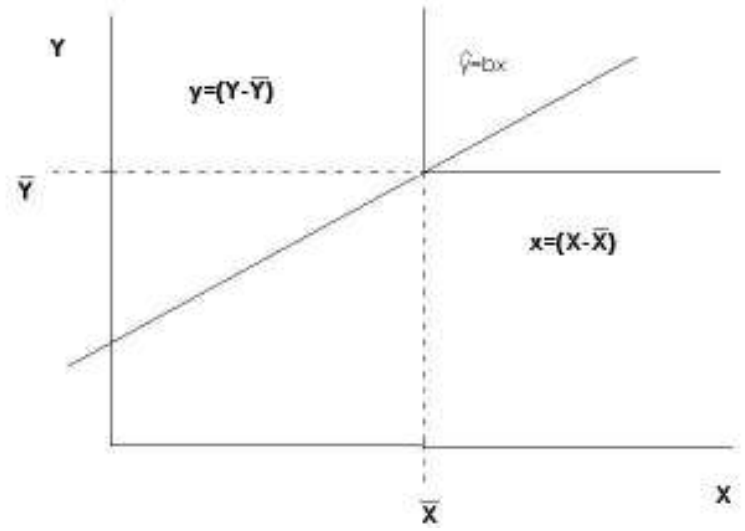
Como puede observarse, dicha ecuación presenta la misma pendiente que la ecuación obtenida en puntuaciones directas. Se diferencia de ésta en que carece de ordenada en el origen. La recta, pues, en centradas pasa por el origen de coordenadas. Esto es:



Obsérvese que las puntuaciones centradas son la consecuencia de restar a los valores Y su media $(Y - \overline{Y})$ y a los valores X, igualmente su media $(X - \overline{X})$. Por otro lado, sucede, precisamente, que tanto la media de Y como la media de X satisfacen la ecuación de la recta, como se desprende de (1.20):

$$\hat{Y} = a + bX$$

Por tanto, la transformación en puntuaciones centradas equivale a un desplazamiento del origen del sistema de coordenadas al punto $(\overline{X},\overline{Y})$. En consecuencia, la recta de regresión observada tendrá la misma pendiente pero carecerá de ordenada en el origen, tal como se observa en la siguiente figura:



c) Puntuaciones estandarizadas

Tomemos como referencia la siguiente ecuación conocida:

$$\hat{Y} - \overline{Y} = b(X - \overline{X})$$

Y sustituyamos *b* por su expresión mínimo cuadrática:

$$\hat{Y} - \overline{Y} = b(X - \overline{X}) = \left(r_{xy} \frac{S_y}{S_x} \right)(X - \overline{X})$$

Se observa que el primer miembro de la igualdad hace referencia a las puntuaciones típicas de *Y,* y el segundo miembro, a las puntuaciones típicas de *X*. Sustituyendo por la notación adecuada:

$$\hat{Z}_y = r_{xy} Z_x$$

Se comprueba que la ecuación en puntuaciones estandarizadas tiene por pendiente el coeficiente de correlación simple.

Ejemplo 1.5.- Sobre los datos del ejemplo 1.1, calcular la ecuación de regresión en puntuaciones directas, centradas y estandarizadas:

SOL:

a) Directas:

$$b = r_{xy}\frac{S_y}{S_x} = 0.8327\frac{2.579}{10.874} = 0.1975$$

$$a = \overline{Y} - b\overline{X} = 6.5 - 0.1975 * 117.5 = -16.702$$

Por tanto:

$$\hat{Y} = a + bX = -16.702 + 0.1975X$$

b) Centradas:

$$\hat{y} = bx = 0.1975x$$

c) Estandarizadas:

$$\hat{Z}_y = r_{xy}Z_x = 0.8327Z_x$$

5.- Interpretación del modelo de regresión

Como se ha indicado, en el modelo de regresión lineal se establece la relación existente entre las variables *X* e *Y*. Esta relación, para todo sujeto, tiene un componente estructural (lineal) de carácter determinista indicado por $a + bX$ y un componente aleatorio *e*, específico para cada individuo. Así:

$$Y = a + bX + e$$

donde la parte determinista que permite obtener la puntuación pronosticada por el modelo es:

$$\hat{Y} = a + bX$$

Distinguimos pues, los siguientes elementos: a) error de estimación -*e*-, b) puntuación pronosticada -$\hat{Y}$-, c) pendiente de la recta -*b*- y d) ordenada en el origen -a-.

a) Error de estimación

La parte aleatoria hace referencia justamente a aquello que el modelo no explica. Muestra la deficiencia del modelo, aunque es obvio que ningún modelo en ciencias humanas, dada su

complejidad, carecer de error. El estudio del error o *puntuaciones residuales* tiene especial interés, como se ver mas adelante en la verificación de los supuestos del modelo. Por el momento, señalemos su existencia. En el ejemplo 1.1, el sujeto número 4, que presenta un coeficiente intelectual -*C.I.*- de 124 puntos, ha obtenido una calificación de 7 puntos. El pronóstico de la ecuación de regresión será:

$$\hat{Y} = a + bX = -16.702 + 0.1975 * 124 = 7.788$$

Y el error obtenido:

$$e = Y - \hat{Y} = 7 - 7.788 = -0.788$$

La interpretación es obvia; para un sujeto de 124 de C.I. el modelo predice 7.788 puntos. Ha obtenido 7 puntos, luego la parte que no explica el modelo corresponde a -0.788 puntos.

b) Puntuación estimada

Mayor interés tiene por el momento que nos concentremos en la parte estructural del modelo. A este respecto hay que decir que el valor $\hat{Y}_i$ obtenido al aplicar la ecuación de regresión sobre un determinado valor X_i hace referencia al *valor promedio* previsto para todos aquellos sujetos que han obtenido en la variable X el valor de X_i . Por ejemplo, en el caso que nos concierne para el sujeto que ha logrado 124 puntos de C.I. la puntuación prevista ha sido de 7.788. Se interpreta como la calificación media de todos los sujetos de 124 puntos en inteligencia. Es obvio que no todos los sujetos de igual inteligencia sacarán exactamente la misma puntuación. Dependiendo de otros factores (motivación, personalidad... etc) unos obtendrán más y otros menos. Al final es el valor más probable (promedio) el especificado por la ecuación de regresión.

c) Pendiente de la recta

La pendiente de la recta tiene una interpretación sencilla en matemáticas; muestra el cambio en Y por cada unidad de cambio en X. Como la ecuación de regresión opera (mediante el procedimiento de mínimos cuadrados) sobre la base del diagrama de dispersión, la interpretación, en este caso, tal como quedo de manifiesto en el apartado anterior, es la siguiente: la pendiente b indica el cambio *medio* en Y asociado a cada unidad de cambio en X. Por ejemplo, en el caso que estamos tratando, la pendiente vale 0.1975. Se interpreta en el sentido de que por cada punto de incremento en el C.I. los sujetos, por término medio, mejorarán en 0.1975 puntos su rendimiento académico.

Una pendiente de cero indica claramente que la variable X no sirve para nada, pero una pendiente grande no indica lo contrario, ya que para esto hace falta conocer las escalas de las variables, y lo que es más importante, la dispersión de la nube de puntos. Un diagrama de dispersión mas bien redondeado, aunque con una recta implícita de gran pendiente no significa gran cosa en términos de relación.

c) Ordenada en el origen

Como se sabe, la ordenada en el origen hace referencia al valor en *Y* cuando *X=0*. En la ecuación de regresión, ya que la recta está elaborada sobre los puntos medios del diagrama de dispersión, hace referencia a la puntuación media de *Y* cuando el valor de *X* es cero. No siempre es interpretable este valor en Psicología. Por ejemplo, en nuestro caso la ordenada en el origen es -16.702. Es evidente que un sujeto no obtendrá esta calificación cuando *X=0*. Los valores negativos en rendimiento carecen de interpretación. Por otro lado, ha de tenerse en cuenta que no es posible encontrar una inteligencia de valor cero; el rango de variación en las variables no ha de estar fuera de los observados en la muestra, ya que éste ha sido el punto de referencia para determinar la ecuación de regresión. Por tanto, aunque la recta pueda prolongarse hasta el infinito no es lícito operar con valores fuera de los márgenes estudiados.

No obstante, frecuentemente, puede interpretarse el valor de la ordenada en el origen. Supongamos que relacionamos la variable *Ingresos* (*Y*) con *Años de estudio* (*X*) y obtenemos la siguiente ecuación de regresión:

$$\hat{Y} = 600 + 120X$$

En este caso, los sujetos que carecen de todo tipo de estudio ganan por término medio 600 euros, de tal manera que por cada año de estudio ven incrementado su salario en 120 euros. Así, un sujeto que haya estudiado 10 años tendrá un sueldo de 600+120*10=1800 euros.

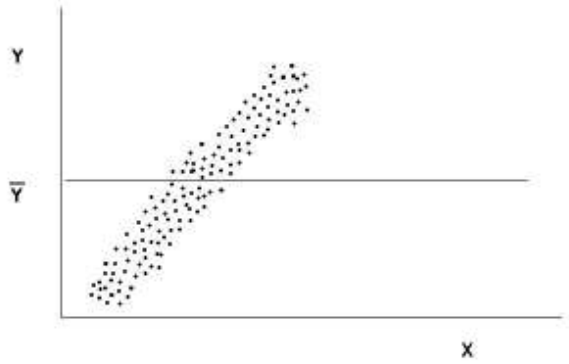
6.- Componentes de variación.

Interesa en este apartado analizar la capacidad predictiva del modelo de regresión lineal. Supuesto que los datos observados se ajustan a una ecuación lineal hemos determinado en el punto anterior aquella recta que mejor cumple dicha condición en el sentido de generar la mínima cantidad de errores cuadráticos posibles. Veremos ahora, en una primera instancia, cuanto, en términos de variación, explica el modelo lineal del conjunto de los datos observados (*bondad de ajuste*) para tratar más adelante de la lógica de la decisión que permite aceptar o rechazar la hipótesis del modelo lineal para un determinado conjunto de datos (*validez del modelo*).

Expondremos, primeramente, los distintos *componentes de variación* que pueden reconocerse al aplicar el modelo regresión sobre un determinado fenómeno observado. Digamos que todo modelo es un intento de explicar la realidad. Y los modelos estadísticos se aplican, precisamente, cuando la realidad estudiada es imperfectamente conocida. Se observa, así, que una parte del comportamiento del fenómeno queda explicado por el modelo, mientras que otra parte se sustrae al mismo.

Para aclarar estas ideas, supongamos en primer lugar que disponemos de dos variables *X* e *Y* pero desconocemos la naturaleza de la relación entre ambas variables. En este supuesto, si nos piden el valor en *Y* para un sujeto que haya obtenido un cierto valor en *X*, daremos como valor más probable la media de *Y*. Es razonable tal respuesta, ya que en ausencia de información para una variable que sigue una ley normal el valor de máxima probabilidad es

precisamente su valor medio. Así pues, como se observa en el siguiente gráfico el valor de Y estimado para cualquier valor de X será $\overline{Y}$:



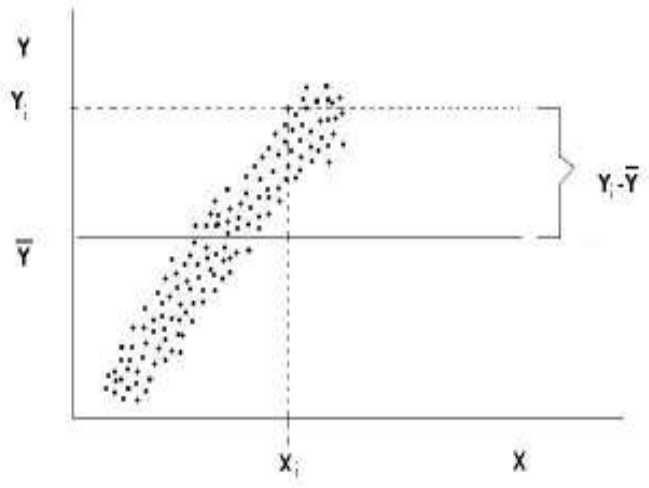
De esta forma, en ausencia total de información, la ecuación de regresión será:

$$\hat{Y} = \overline{Y}$$

Para un sujeto en particular que dado un valor X_i haya obtenido Y_i, cometeremos un error de predicción:

$$e = Y_i - \overline{Y}$$

tal como se ilustra en la siguiente figura:



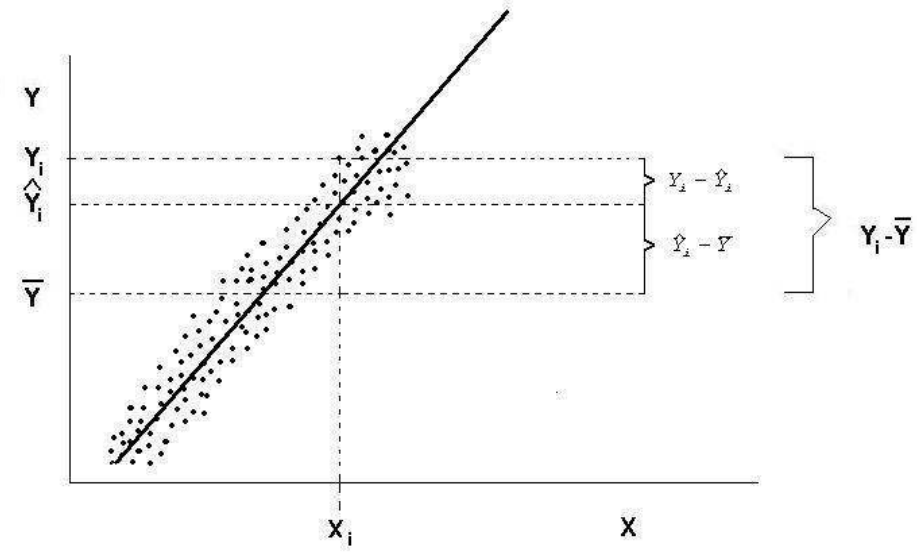
Supongamos ahora que tenemos conocimiento de la relación lineal que liga las variable X e Y. Y esta relación es según la ecuación conocida

$$\hat{Y} = a + bX$$

El error cometido será entonces:

$$Y_i - \hat{Y}_i$$

según se ilustra en la siguiente figura:



Se observa que en este caso el error es más pequeño que el existente en ausencia de información. Si tomamos el valor:

$$Y_i - \bar{Y}$$

como indicativo del error cometido cuando carecemos de la información proporcionada por el modelo y lo definimos como desviación total respecto a la media para un determinado sujeto, entonces el valor:

$$\hat{Y}_i - \bar{Y}$$

hará referencia a la parte que de la desviación total explica el modelo de regresión. Se denomina *desviación explicada* por el modelo de regresión. Queda, entonces, un resto:

$$Y_i - \hat{Y}_i$$

que no logra explicar el modelo *-desviación no explicada-*. De esta forma, según lo expuesto, podemos establecer la siguiente igualdad:

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y})$$

Para el sujeto *i* la *desviación total* del valor Y_i con respecto a la media $(Y_i - \overline{Y})$puede descomponerse en la *desviación explicada* por el modelo de regresión $(\hat{Y}_i - \overline{Y})$ más la *desviación no explicada* $(Y_i - \hat{Y})$.

Si elevamos al cuadrado ambos miembros de la igualdad (1.30):

$$(Y_i - \overline{Y})^2 = (\hat{Y}_i - \overline{Y})^2 + (Y_i - \hat{Y})^2 + 2(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y})$$

Si se cumple esta igualdad para cada uno de los sujetos, se cumplirá igualmente para la suma de todos ellos. Así pues:

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y})^2 - 2\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y})$$

Donde:

$$2\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}) = 0$$

ya que los errores aleatorios no correlacionan con ninguna otra puntuación (Obsérvese que el sumatorio anterior es el numerador de la covarianza entre los errores y las puntuaciones predichas por el modelo de regresión). En consecuencia:

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y})^2$$

Esto es:

Suma de cuadrados total = Suma de cuadrados explicada + Suma de cuadrados no explicada

Si tomamos las sumas de cuadrados anteriores (como numeradores de varianzas que son) como un índice de la variabilidad de los datos tenemos que:

Variación Total = Variación Explicada + Variación No Explicada.

Merece destacarse la importancia de esta igualdad. Del cociente entre la variación explicada y

la total obtendremos la proporción de variación explicada por el modelo, lo que permitirá hacernos una idea del ajuste del modelo al fenómeno observado -*bondad de ajuste*-. Por otro lado, a partir de estos datos calcularemos la *varianza explicada* y *no explicada*, permitiéndonos su cociente tomar la decisión de si el modelo lineal es un buen indicador del comportamiento de los datos observados -*validez del modelo*-. A estas consideraciones dedicamos los dos próximos apartados.

7.- Bondad de ajuste

Tomaremos como índice de la bondad de ajuste del modelo la proporción de variación explicada por el mismo; esto es, el cociente entre la suma de cuadrados explicada por el modelo y la suma de cuadrados total. De esta forma, podemos hacernos una idea de cuánto explica el modelo de la realidad estudiada. Su expresión es:

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}$$

Obsérvese que este cociente lo hemos denominado como R^2. Coincide, precisamente, como demostraremos a continuación con el valor de r_{xy} al cuadrado, también denominado *coeficiente de determinación*. En este sentido, en relación al numerador de la expresión (1.32) se sabe que la ecuación de regresión en puntuaciones centradas es:

$$\hat{Y}_i - \overline{Y} = b(X_i - \overline{X})$$

Elevando al cuadrado y sacando sumatorios:

$$\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2 = b^2 \sum_{i=1}^{N} (X_i - \overline{X})^2$$

Por otro lado, se sabe que $\sum_{i=1}^{N} (X_i - \overline{X})^2$ representa el numerador de la varianza de X. Así pues:

$$\sum_{i=1}^{N} (X_i - \overline{X})^2 = NS_x^2$$

Igualmente, en relación a $\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2$:

$$\sum_{i=1}^{N} (Y_i - \overline{Y})^2 = NS_y^2$$

Por tanto, podremos expresar (1.32) de la siguiente manera:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \frac{b^2 N S_x^2}{N S_y^2} = \frac{b^2 S_x^2}{S_y^2}$$

Sustituyendo *b* por su expresión mínimo cuadrática:

$$R^2 = \frac{b^2 S_x^2}{S_y^2} = \frac{\left(r_{xy}\frac{S_y}{S_x}\right)^2 S_x^2}{S_y^2} = r_{xy}^2$$

De donde se comprueba que la proporción de varianza explicada corresponde con el valor de r_{xy}^2. De esta forma, la interpretación de R^2 es extremadamente sencilla y clarificadora. En el ejemplo 1.1 se obtuvo $r_{xy} = 0.8327$. Por tanto, el cuadrado de este valor, $R^2 = 0.8327^2 = 0.6933$ nos indica que el 69.33% de la variación observada en el rendimiento es debida a la inteligencia. Queda, en consecuencia, un 30.67% de variación debido a otros factores (motivación, horas de estudio, ..etc).

Resulta patente, pues, la utilidad de R^2 para hacernos una ideal cabal del efecto de una variable sobre otra. En términos prácticos, para calcular la bondad de ajuste del modelo bastar con elevar al cuadrado el coeficiente de correlación (r_{xy}) que se supone ya ha sido obtenido en su momento (ver fórmula (1.9) o equivalente). También podemos aplicar la fórmula (1.33), si disponemos de las varianzas de *X* y de *Y*. Podemos, igualmente, aplicar directamente la expresión (1.32) o bien, si operamos en base a las puntuaciones directas utilizaremos la siguiente:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \frac{b^2\sum_{i=1}^{N}(X_i - \overline{X})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \frac{b^2\left(\sum_{i=1}^{N}X_i^2 - \frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)}{\sum_{i=1}^{N}Y_i^2 - \frac{\left(\sum_{i=1}^{N}Y_i\right)^2}{N}}$$

Por otro lado, podemos replantear la fórmula (1.31) en función de R^2. De esta forma logramos una mejor comprensión de dicha igualdad, al mismo tiempo que al expresarse en términos de proporción quedamos liberados de los problemas de las escalas. Para ello dividamos los dos miembros de la igualdad (1.31) por la suma de cuadrados total:

$$\frac{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}=\frac{\sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2}{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}+\frac{\sum_{i=1}^{N}(Y_i-\hat{Y})^2}{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}$$

Esto es:

Prop. variabilidad total = prop. variabilidad explicada + prop. variabilidad no explicada

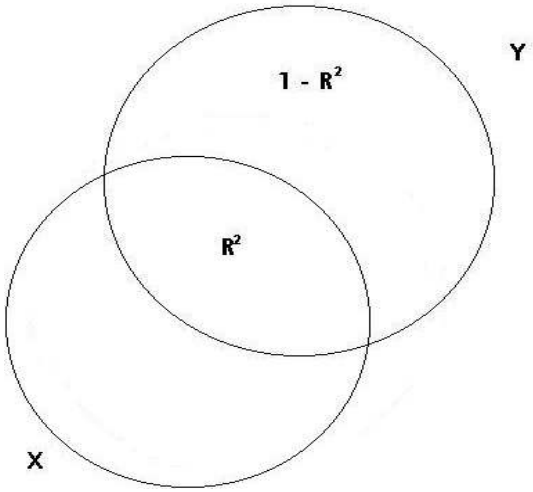
Es fácil deducir que:

$$\textit{Prop. var. no explicada} = 1 - R^2$$

Luego la expresión (1.35) deviene:

$$1 = R^2 + (1 - R^2)$$

Gráficamente el reparto de variabilidad podemos representarlo en el siguiente diagrama de Venn. La intersección de los círculos indica la proporción de variabilidad explicada por la regresión:



Ejemplo 1.6.- Determinar los componentes de variación y la proporción de variación explicada por el modelo de regresión lineal de los datos del ejemplo 1.1.

SOL:

Comenzaremos con la expresión original (1.32), que no es precisamente la fórmula más simple de realizar, pero tiene la ventaja de ser la que mejor refleja la lógica de la bondad de ajuste. Permite distinguir para cada puntuación de *Y* los distintos componentes de variación (desviación explicada, no explicada y total):

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}$$

Calculemos, en primer lugar, la suma de cuadrados total:

$$\sum_{i=1}^{N} (Y_i - \overline{Y})^2 = (4-6.5)^2 + (8-6.5)^2 + (2-6.5)^2 + (7-6.5)^2 + (9-6.5)^2 +$$
$$+ (9-6.5)^2 + (3-6.5)^2 + (10-6.5)^2 + (7-6.5)^2 + (6-6.5)^2 = 66.5$$

Antes de proceder a calcular la suma de cuadrados explicada, hemos de determinar los valores predichos por la ecuación de regresión para los distintos valores de *X*. Así pues:

$$\hat{Y}_1 = -16.702 + 0.1975 * 105 = 4.032$$
$$\hat{Y}_2 = -16.702 + 0.1975 * 116 = 6.204$$
$$\hat{Y}_3 = -16.702 + 0.1975 * 103 = 3.637$$
$$\hat{Y}_4 = -16.702 + 0.1975 * 124 = 7.784$$
$$\hat{Y}_5 = -16.702 + 0.1975 * 137 = 10.351$$
$$\hat{Y}_6 = -16.702 + 0.1975 * 127 = 8.178$$
$$\hat{Y}_7 = -16.702 + 0.1975 * 112 = 5.414$$
$$\hat{Y}_8 = -16.702 + 0.1975 * 129 = 8.771$$
$$\hat{Y}_9 = -16.702 + 0.1975 * 118 = 6.599$$
$$\hat{Y}_{10} = -16.702 + 0.1975 * 105 = 4.032$$

Una vez obtenidas las puntuaciones estimadas por el modelo procedemos a calcular la suma de cuadrados explicada:

$$\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2 = (4.032-6.5)^2 + (6.204-6.5)^2 + (3.637-6.5)^2 + (7.784-6.5)^2 + (10.351-6.5)^2 +$$
$$+ (8.178-6.5)^2 + (5.414-6.5)^2 + (8.771-6.5)^2 + (6.599-6.5)^2 + (4.032-6.5)^2 = 46.108$$

De aquí se deduce que la suma de cuadrados no explicada será:

$$\sum_{i=1}^{N} (Y_i - \hat{Y})^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 + \sum_{i=1}^{N} (\hat{Y}_i - \hat{Y})^2 = 66.5 - 46.108 = 20.392$$

Y la proporción de variabilidad explicada:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \frac{46.108}{66.5} = 0.6933$$

Otra fórmula más útil para el calculo de R^2 es:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \frac{b^2\left(\sum_{i=1}^{N}X_i^2 - \frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)}{\sum_{i=1}^{N}Y_i^2 - \frac{\left(\sum_{i=1}^{N}Y_i\right)^2}{N}} = \frac{0.1975^2\left(139245 - \frac{1175^2}{10}\right)}{489 - \frac{65^2}{10}} = \frac{46.108}{66.5} = 0.6933$$

O más sencilla aún:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} = \frac{b^2 N S_x^2}{N S_y^2} = \frac{0.1975^2 * 10 * 10.874^2}{10 * 2.579^2} = \frac{46.108}{66.5} = 0.6933$$

8.-- Validación del modelo.

Como se ha indicado, hay dos de variación en todo fenómeno de base estadística: la fuente de variación especificada por el modelo y que constituye su *estructura*, y una fuente de variación aleatoria, no controlada, que imprime una cierta deformación sobre el modelo concebido. Desde esta perspectiva, la validación del modelo consiste básicamente en comprobar si persiste la estructura del modelo a pesar de la deformación a por la fluctuación aleatoria de los datos.

A nivel estadístico, se trata de comparar la *varianza explicada*, que define el modelo, con la *varianza no explicada*, que lo desdibuja. Si la varianza explicada es mayor que la no explicada será indicativo de que se reconoce algo a pesar del ruido, si ocurre lo contrario, el ruido, la deformación que impone la varianza aleatoria impedirá toda posibilidad de reconocimiento y el modelo no será validado.

La prueba estadística que permite comparar varianzas y tomar decisiones en cuanto a su

magnitud relativa es, como se sabe, el *análisis de la varianza*. A dicha prueba nos remitimos cuando hablamos de validación del modelo.

A este respecto, la varianza explicada tendrá por valor:

$$\frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{k}$$

Siendo el numerador la suma de cuadrados explicada por la regresión y el denominador los grados de libertad asociados al componente de variación explicado, donde k indica el número de variables independientes a considerar.

Por otro lado, la varianza no explicada será:

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{N-k-1}$$

donde el numerador hace referencia a la suma de cuadrado no explicada por el modelo, y el denominador sus grados de libertad asociados (*N* hace referencia al número de individuos y *k* al número de variables independientes).

El análisis de la varianza queda, entonces, de la siguiente manera:

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{k}}{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{N-k-1}}$$

Si el valor obtenido de *F* es superior al de las tablas para *k* y *N-k-1* grados de libertad y al nivel de significación de α, rechazaremos la hipótesis de igualdad de varianzas (con un riesgo máximo α). Concluiremos, en consecuencia, que muy probablemente las variables *X* e *Y* están relacionadas. Así:

$$F > F_{(k,N-k-1,\alpha)} \quad \Rightarrow \quad Se\ rechaza\ la\ H_0$$

En caso contrario, si el valor obtenido de *F* es igual o inferior al de las tablas, concluiremos (con un riesgo β desconocido) que ambas varianzas son iguales, y por tanto, no estaremos en condiciones de rechazar la H_0. Concluiremos, por tanto, que muy probablemente las variables *X* e *Y* no estarán relacionadas. Esto es:

$$F \leq F_{(k,N-k-1,\alpha)} \quad \Rightarrow \quad Se\ acepta\ la\ H_0$$

Aunque la fórmula (1.40) es suficiente para determinar la validez del modelo, habitualmente se recurre a la siguiente tabla donde quedan desglosados los distintos elementos que configuran dicha fórmula. De esta forma se ve de una manera más clara los componentes de variación del modelo así como sus grados de libertad asociados.

FUENTE DE VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	VARIANZA	F
Explicada	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	k	$\frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{k}$	$F = \frac{Var_{\text{exp.}}}{Var_{no\text{exp.}}}$
No explicada	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$N-k-1$	$\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{N-k-1}$	
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2$	$N-1$	$\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{N-1}$	

Podemos aplicar la fórmula (1.40) directamente o bien utilizar alguna fórmula alternativa más sencilla. De esta forma, en relación a la suma de cuadrados debida a la regresión podemos utilizar la expresión conocida:

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = b^2\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Si disponemos de las puntuaciones directas de la variable X, resulta más simple:

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = b^2\sum_{i=1}^{n}(X_i - \overline{X})^2 = b^2\left(\sum_{i=1}^{N}X_i^2 - \frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)$$

Más fácil aún, si conocemos la varianza de X, que se supone ha sido calculada previamente:

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = b^2\sum_{i=1}^{n}(X_i - \overline{X})^2 = b^2NS_x^2$$

Ya que:

$$\frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N}=S_x^2 \quad \Rightarrow \quad \sum_{i=1}^{N}(X_i-\overline{X})^2=NS_x^2$$

Y en relación a la suma de cuadrados no explicada (o residual), ésta puede expresarse como la diferencia entre la suma de cuadrados total y explicada:

$$\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{N}(Y_i-\overline{Y})^2-\sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2$$

Y de una manera más sencilla en base a lo expuesto anteriormente:

$$\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{N}(Y_i-\overline{Y})^2-\sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2=NS_y^2-b^2NS_x^2$$

Según utilicemos una u otra expresión tendremos diferentes alternativas a la fórmula (1.40). Por ejemplo, si operamos en puntuaciones directas:

$$F=\frac{\frac{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}{k}}{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{N-k-1}}=\frac{\frac{b^2\left(\sum_{i=1}^{N}X_i^2-\frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)}{k}}{\frac{\sum_{i=1}^{N}Y_i^2-\frac{\left(\sum_{i=1}^{N}Y_i\right)^2}{N}-b^2\left(\sum_{i=1}^{N}X_i^2-\frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)}{N-k-1}}$$

O bien en términos de varianzas, si éstas se conocen:

$$F=\frac{\frac{\sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2}{k}}{\frac{\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2}{N-k-1}}=\frac{\frac{b^2NS_x^2}{k}}{\frac{NS_y^2-b^2NS_x^2}{N-k-1}}$$

Podemos simplificar aún más el cálculo de F, y expresarlo en términos de R^2 según la siguiente fórmula:

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{N-k-1}}$$

Para ello, tan sólo tenemos que dividir el numerador y el denominado de (1.42) por la suma de cuadrados de Y. Así pues:

$$F = \frac{\frac{\sum_{i=1}^{N}(\hat{Y}-\overline{Y})^2}{k}}{\frac{\sum_{i=1}^{N}(Y-\hat{Y})^2}{N-k-1}} = \frac{\frac{\sum_{i=1}^{N}(\hat{Y}-\overline{Y})^2 / \sum_{i=1}^{N}(Y-\overline{Y})^2}{k}}{\frac{\sum_{i=1}^{N}(Y-\hat{Y})^2 / \sum_{i=1}^{N}(Y-\overline{Y})^2}{N-k-1}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{N-k-1}}$$

Ejemplo 1.7.- Calcular la validez del modelo de regresión lineal del ejemplo 1.1.

SOL:

Si lo hacemos en términos de las puntuaciones directas:

$$F = \frac{\frac{b^2\left(\sum_{i=1}^{N}X_i^2 - \frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)}{k}}{\frac{\sum_{i=1}^{N}Y_i^2 - \frac{\left(\sum_{i=1}^{N}Y_i\right)^2}{N} - b^2\left(\sum_{i=1}^{N}X_i^2 - \frac{\left(\sum_{i=1}^{N}X_i\right)^2}{N}\right)}{N-k-1}} = \frac{\frac{0.1975^2\left(139245 - \frac{1175^2}{10}\right)}{1}}{\frac{489 - \frac{65^2}{10} - 0.1975^2\left(139245 - \frac{1175^2}{10}\right)}{10-1-1}} = \frac{\frac{46.108}{1}}{\frac{20.392}{8}} = 18.088$$

Buscando en las tablas:

$$F_{(1,8,0.05)} = 5.318$$

Comparando:

$$18.088 > 5.318$$

Luego se rechaza la H_0 (con un riesgo máximo de 0.05). Puede considerarse válido el modelo.

Si operamos en términos de varianzas:

$$F = \frac{\frac{b^2 NS_x^2}{k}}{\frac{NS_y^2 - b^2 NS_x^2}{N-k-1}} = \frac{\frac{0.1975^2 * 10 * 10.874^2}{1}}{\frac{10 * 2.579^2 - 0.1975^2 * 10 * 10.874^2}{8}} = 18.088$$

Más fácilmente podemos aplicar la expresión (1.43) para el cálculo de la validez. Así:

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{N-k-1}} = \frac{\frac{0.8237^2}{1}}{\frac{1-0.8237^2}{8}} = 18.088$$

Si se desea, a efectos ilustrativos de los distintos elementos que componen el análisis de la varianza, podremos elaborar la siguiente tabla:

FUENTE DE VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	VARIANZA	F
Explicada	46.108	1	46.108	$F = 18.088$
No explicada	20.392	8	2.549	
Total	66.5	9	7.389	

9.- Significación de los parámetros de la regresión.

La significación de los parámetros del modelo de regresión reviste especial interés en el contexto de la regresión múltiple, donde pudiera ocurrir que la prueba *F* del análisis de la varianza mostrara que en términos globales el modelo fuera válido, mientras que el efecto de algunas variables del modelo fuera nulo, o lo que es lo mismo que algunos coeficientes de regresión no ejercieran ningún efecto significativo sobre la variable dependiente.

En el caso de la regresión simple -ya que existe una sola variable independiente- la prueba de significación de los coeficientes de regresión puede considerarse como una prueba equivalente a la prueba del análisis de la varianza (también de la significación del coeficiente de correlación r_{xy}).

De los dos coeficientes de regresión del modelo (a y b) nos interesan tan sólo la pendiente de la recta, que es precisamente el coeficiente que nos muestra el efecto de la variable *X* sobre *Y*. En concreto comprobaremos si su valor es estadísticamente igual a cero o no. Si dicha pendiente no difiere significativamente de cero concluiremos que el modelo no aporta información relevante. En caso contrario, daremos el modelo como válido. Esto es, tengamos la ecuación de regresión en puntuaciones centradas:

$$\hat{Y} - \overline{Y} = b(X - \overline{X})$$

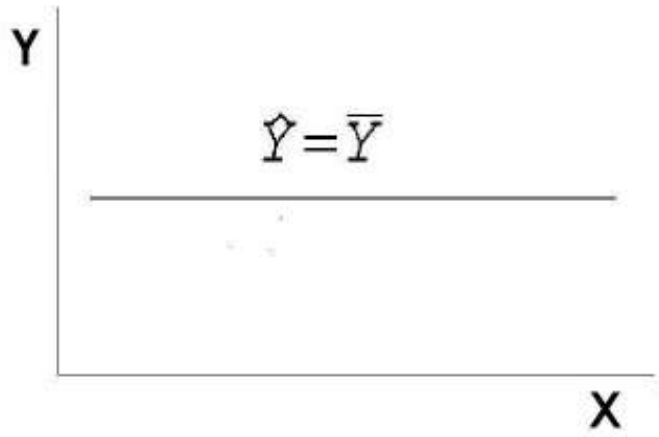
Despejando Y:

$$\hat{Y} = \overline{Y} + b(X - \overline{X})$$

Se observa que cuando la pendiente vale cero:

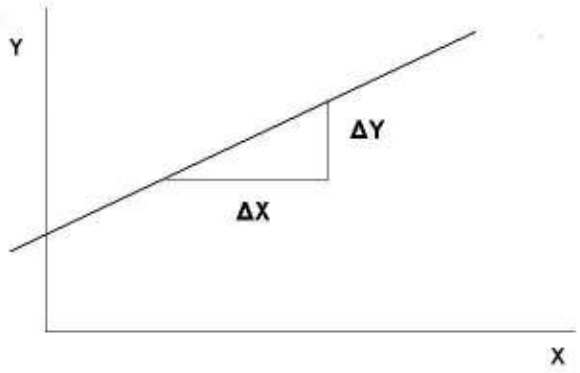
$$\hat{Y} = \overline{Y}$$

la puntuación pronosticada es precisamente la media de Y, (que, como se sabe, es el pronóstico cuando el modelo carece de información alguna) tal como se refleja en el siguiente gráfico:



La recta de regresión es pues, paralela al eje de las abscisas. Cualquier cambio en X implica siempre el mismo valor en Y. Ambas variables no están, por tanto, correlacionadas.

Por otro lado, como es fácil comprobar, cuando la pendiente de la recta es diferente de cero incrementos en el valor de *X* suponen incrementos efectivos en la variable *Y*:



En esta situación las variables están relacionadas, el modelo aporta información relevante en términos predictivos y es, por ello, válido.

Así pues, como se ha indicado, la validez del modelo puede comprobarse también (además de la prueba *F*) contrastando la pendiente asociada al modelo de regresión. Si se demuestra que la pendiente es significativamente diferente de cero, el modelo tendrá capacidad predictiva, y por tanto, será válido. Por el contrario, si la pendiente no fuera

estadísticamente diferente de cero su capacidad predictiva no irá más allá de $\overline{Y}$ (predicción en ausencia de información) y el modelo no será válido.

En términos estadísticos se trata de comprobar si la pendiente *b* observada en una cierta muestra puede o no proceder de una población cuya pendiente β vale cero. Esto es, se contrasta la hipótesis nula:

$$H_0 : \beta = 0$$

frente a la hipótesis alternativa:

$$H_1 : \beta \neq 0$$

En este supuesto, se demuestra (ver Apéndice A) que la distribución muestral de coeficientes b procedentes de una población cuyo valor es cero, se distribuye según una ley de *Student* de media cero y desviación tipo:

$$S_{b_i} = \sqrt{\frac{S_{res}^2}{\sum_{i=1}^{n}(X - \overline{X})^2}} = \sqrt{\frac{S_{res}^2}{NS_x^2}}$$

De esta forma, si se desea saber si un determinado coeficiente b observado en una muestra procede de una población de $\beta = 0$, calcularemos el número de desviaciones tipo que se encuentra de la media de dicha distribución, según la fórmula conocida:

$$t = \frac{b - \beta}{S_{b_i}} = \frac{b - 0}{\sqrt{\frac{S_{res}^2}{\sum_{i=1}^{n}(X - \overline{X})^2}}}$$

Posteriormente comparamos este valor *t* con el de las tablas $t_{(\alpha, N-2)}$ para el nivel de significación α y *N-2* grados de libertad:

Si $t \leq t_{(\alpha, N-2)}$ Se acepta la hipótesis nula. El modelo no es válido

Si $t > t_{(\alpha, N-2)}$ Se rechaza la hipótesis nula. El modelo es válido

Ejemplo 1.8.- Determinar la significación del coeficiente de regresión de ejemplo 1.3.

SOL:

Apliquemos (1.45):

$$t = \frac{b - 0}{\sqrt{\frac{S_{res}^2}{\sum_{i=1}^{n}(X - \overline{X})^2}}} = \frac{0.1975}{\sqrt{\frac{2.549}{1182.5}}} = 4.253$$

Buscamos la t de las tablas para $\alpha = 0.05$ y $N - 2 = 8$ grados de libertad:

$$t_{(0.05,8)} = 2.306$$

Comparándolo con el valor obtenido:

$$4.253 > 2.306$$

La pendiente es significativamente distinta de cero. Existe, pues, relación entre ambas variables.

10.- Predicción.

Una vez validado el modelo de regresión que liga las variables X e Y puede ser conveniente utilizarlo para establecer predicciones de la variable Y. Por ejemplo, si conocemos para una cierta muestra de vendedores la relación existente entre una determinada prueba psicológica y el éxito profesional de los mismos, puede interesarnos, si disponemos de un candidato a vendedor, aplicar dicha prueba a efectos de su capacidad en ventas.

Si para la elaboración del modelo dispusiéramos de los datos de toda la población sucedería que la ecuación de regresión obtenida sería precisamente la ecuación regresión verdadera

$$\hat{Y} = \alpha + \beta X$$

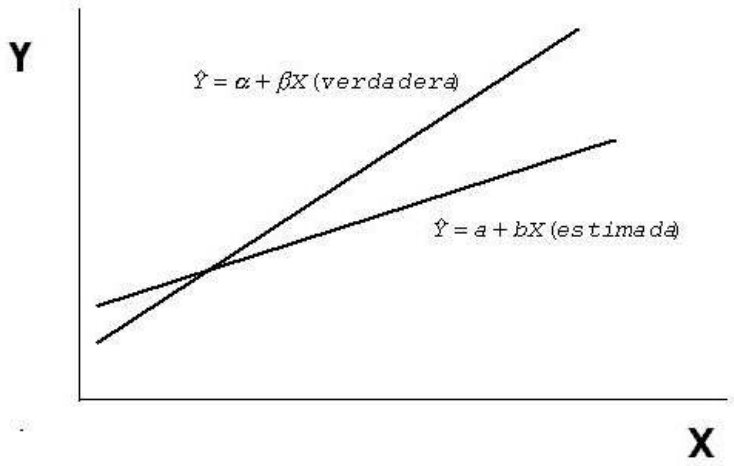
En este supuesto, el valor más probable en *Y* para un sujeto que haya obtenido un cierto valor en *X* sería el reflejado en la ecuación de regresión (1.46):

Si deseamos afinar algo más y ofrecer una estimación por intervalo, sabemos por los requisitos del modelo de regresión que para un cierto valor X_o la distribución ligada de los valores Y sigue una ley normal de media el valor predicho en la ecuación de regresión y de varianza la varianza residual. De esta forma, para los sujetos que han obtenido X_o habrá una proporción $1-\alpha$ de ellos que tendrán en Y puntuaciones comprendidas en el siguiente intervalo:

$$\hat{Y}_0 \pm t_{(N-2,\alpha)} S_e$$

En términos de probabilidad, diremos que un sujeto que ha obtenido una cierta puntuación X_o tendrá una probabilidad $1-\alpha$ de estar comprendido en los citados límites.

En la práctica, no obstante, sucede que desconocemos la recta de regresión verdadera; tan sólo disponemos de la recta de regresión obtenida en una muestra. En consecuencia, entre la ecuación de regresión estimada y la verdadera habrá una cierta diferencia tal como se muestra en la siguiente figura:



No podemos especificar el valor exacto del error ya que desconocemos los parámetros poblacionales. Lo que sí podemos cuantificar es la distribución en el muestreo de los distintos valores $\hat{Y}_o$ en torno al valor real Y_o. Esto es, hemos de determinar la Var(Y_o). A este respecto, se sabe que:

$$Y_0 = a + bX_0 + e$$

Luego:

$$Var(Y_0) = Var(a + bX_0 + e) = Var\,(a) + X_0^2 Var\,(b) + S_{res}^2$$

Pero sabemos (ver Apéndice A) que:

$$Var(a) = \left(\frac{1}{N} + \frac{\overline{X}^2}{\sum_{i=1}^{N} (X - \overline{X})^2} \right) S_{res}^2$$

En consecuencia:

$$Var\,(Y_0) = \left(\frac{1}{N} + \frac{\overline{X}^2}{\sum_{i=1}^{N} (X - \overline{X})^2} \right) S_{res}^2 + X_0^2 \frac{S_{res}^2}{\sum_{i=1}^{N} (X - \overline{X})^2} + S_{res}^2$$

Haciendo operaciones:

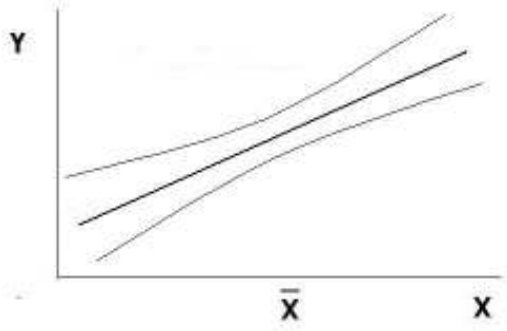
$$Var\,(Y_0) = S_{res}^2 \left(1 + \frac{1}{N} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{N} (X - \overline{X})^2} \right)$$

Por tanto, el intervalo de confianza será:

$$\hat{Y}_0 \pm t_{(N-2,\alpha)} \sqrt{S_{res}^2 \left(1 + \frac{1}{N} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{N} (X - \overline{X})^2} \right)}$$

siendo X_0 es el valor especificado de X sobre el que se desea la predicción. Obsérvese cómo cuanto más alejado se encuentre este valor de la media mayor dispersión habrá para el intervalo de confianza de la $\hat{Y}_0$.

En el siguiente gráfico se muestra dos líneas ligeramente curvas que indican las distintas amplitudes de los intervalos de confianza a lo largo del recorrido de la ecuación de regresión. Tales amplitudes son menores cuanto más cerca se encuentre de la media $\overline{X}$:



Ejemplo 1.9.- Tomando como referencia los datos del ejemplo 1.3, determinar el la calificación verdadera para una persona que presenta 115 puntos de C.I.

SOL:

Aplicando la ecuación de regresión tenemos que la puntuación pronosticada para este sujeto será:

$$\hat{Y}_0 = a + bX = -16.702 + 0.1975 * 115 = 6.011$$

Y el intervalo de confianza ($\alpha = 0.05$) donde espera encontrarse el parámetro correspondiente:

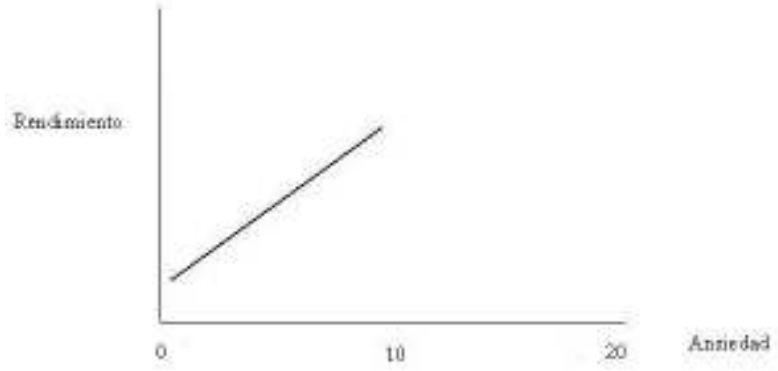
$$\hat{Y}_0 \pm t_{(N-2,\alpha)} \sqrt{S^2_{res}\left(1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{N}(X - \bar{X})^2}\right)} = 6.011 \pm 2.306 \sqrt{2.549\left(1 + \frac{1}{10} + \frac{(115 - 117.5)^2}{1182.5}\right)} =$$

$$6.011 \pm 3.871 = 2.140 \Leftrightarrow 9.882$$

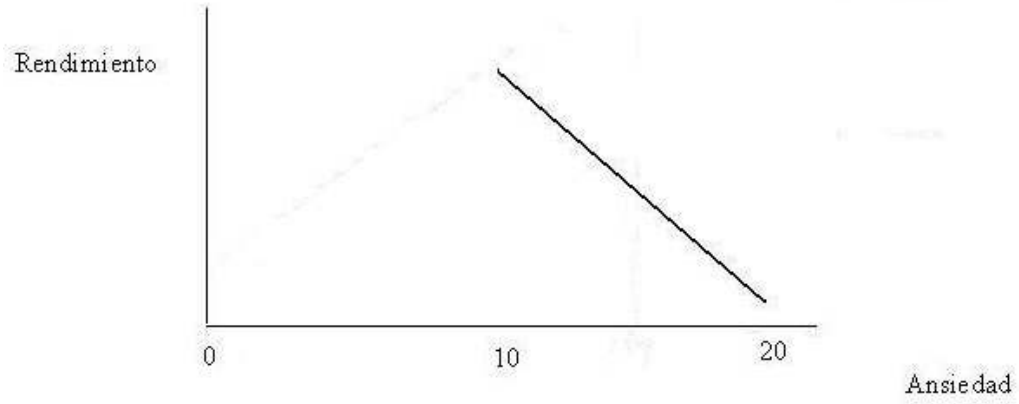
Existe una probabilidad de 0.95 de que un sujeto que presente un C.I. de 115 obtenga entre 9.882 y 2.140 en rendimiento. Obsérvese la magnitud del intervalo que hace posible prácticamente cualquier calificación (de suspenso a sobresaliente) debido a la muestra tan pequeña (10 sujetos) que por motivos didácticos ha sido utilizada.

10.1.- Limitaciones de la predicción

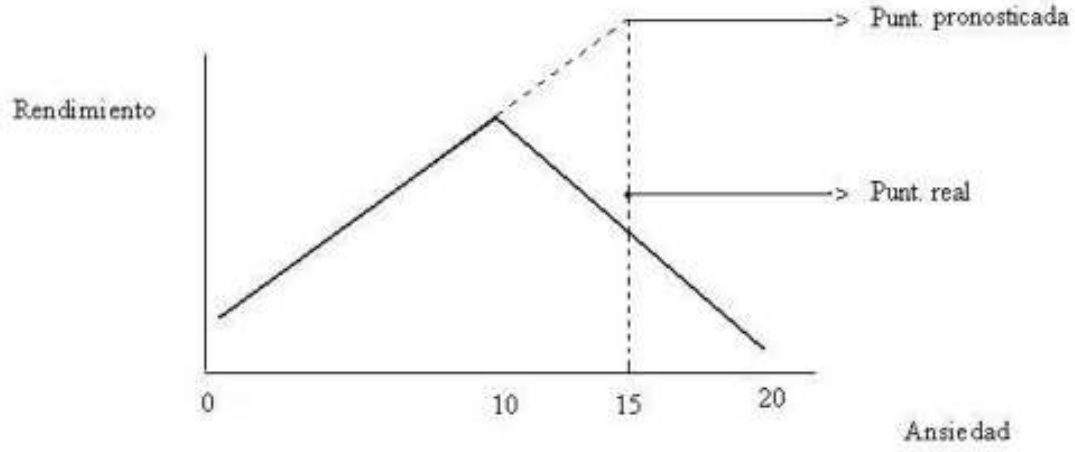
Es preciso hacer algunas consideraciones en relación al alcance de la predicción Aunque la ecuación de la recta puede prolongarse indefinidamente en sus dos extremos, hay que tener la precaución de no extrapolar los valores más allá de los datos de observación. La ecuación de regresión ha sido obtenida a partir de unos determinados valores muestrales, y a estos valores hay que atenerse. Pudiera ocurrir que dentro del rango de observación existiese una relación lineal, pero al mismo tiempo, fuera de ese rango la linealidad dejara de existir. Por ejemplo, supongamos que estudiamos el efecto de la ansiedad sobre el rendimiento dentro de un rango de 0 a 10 en niveles de ansiedad. Podríamos obtener un gráfico de las siguientes características:



Si a continuación estudiásemos esta misma relación pero para un rango en ansiedad de 10 a 20 puntos, imaginemos que la relación fuera de este tipo:



Supongamos que un determinado investigador que desconoce el segundo estudio trabaja con un sujeto cuyo nivel de ansiedad es de 15 puntos. En base a lo que conoce del tema se sentirá inclinado a extrapolar los valores según el siguiente gráfico:



El error ha sido considerable. Ha supuesto que el rendimiento aumentaba cuando en realidad ha disminuido. De aquí se deduce que hemos de operar con suma precaución a la hora de realizar predicciones estadística y limitarnos siempre al rango de valores sobre los que se ha elaborado el modelo, ya que no tenemos información de lo que ocurre fuera de los límites observados, y pudiera ocurrir que la linealidad quedara desvirtuada fuera de tales márgenes.

Bibliografía.

ACHEN, C. H. (1982). *Interpreting and using regression*. London: Sage.

AIKEN, L., AND WEST, S. (1991). *Multiple regression: Testing and interpreting Interactions.* London: Sage

AMON, J. (1990). *Estadística para psicólogos (1)*. Estadística Descriptiva. Madrid: Pirámide.

AMON, J. (1990). *Estadística para psicólogos (2). Probabilidad. Estadística Inferencial*. Madrid: Pirámide.

BOTELLA Y SANMARTIN, R. (1992). *Análisis de datos en Psicología I*. Madrid: Pirámide.

BOTELLA, J. y BARRIOPEDRO, M. I. (1991). *Problemas y ejercicios de Psicoestadística.* Madrid: Pirámide.

BRETT, J. M.; JAMES, L. R. (1982) *Causal Analysis: assumptioms, models and data.* Bervely Hills: SAGE.

COHEN, J. and COHEN, P. (1975). *Applied Multiple Regresion/Correlation analysis for the Behavioral Sciences.* Hillsdales, N. J.: LEA

COOK, R. D. and WEISBERG S. (1982). *Residual and influence in regression. New York*: Chapman & Hall.

CHATTERJEE, S. (1977). *Regression analysis by example*. New York: Wiley.

DOMENECH, J. M. (1985). *Métodos estadísticos: modelo lineal de regresión*. Barcelona: Herder.

DRAPER, N. R. (1986). *Applied regression analysis*. New York: John Wiley

JACCARD, J., LEE TEITEL, TURRISI, R., WAN, C. (1990). *Interaction effects in multiple regression*. Sage University Paper series on Quantitative Applications in the Social Sciences. Newbury Park, CA:Sage

JAMES, L. R. (1982). *Causal analysis: assumptions, models and data*. Bervely Hills: Sage.

JAÑEZ, L. (1980). *Fundamentos de psicología matemática*. Madrid: universidad Complutense.

LEWIS-BECK, M. S. (1980). *Applied regression*. London: Sage.

PEDHAZUR, E. J. (1982). *Multiple regression in behavioral research. Explanation and prediction* (2nd ed.). New York: Halt, Rinehart and Winston.

PEÑA, D. (1987).:*Estadística, modelos y métodos. 2. Modelos lineales y series temporales* Alianza Universidad.

SHOEDER et al. (1982). *Understanding regression analysis: an introductory guide.* Bervely Hills: Sage.

WONNACOTT, T. H. and WONNACOTT, R. J. (1981). *Regression: a second course in statistics.* New York: Wiley.

Internet

Universidad de Cádiz: http://www2.uca.es/serv/ai/formacion/spss/Pantalla/18reglin.pdf
Universidad de California: http://www.ats.ucla.edu/stat/spss/topics/regression.htm
Linear Regression (and Best Fit): http://www.mste.uiuc.edu/patel/amar430/intro.html
Regression analysis: http://en.wikipedia.org/wiki/Regression_analysis
Regression analysis: http://elsa.berkeley.edu/sst/regression.html
Regression to the Mean: http://www.socialresearchmethods.net/kb/regrmean.php
El modelo de regresión lineal simple: http://www.udc.es/dep/mate/estadistica2/cap6.html
Página de Karl Wünsch sobre correlación: http://core.ecu.edu/psyc/wuenschk/docs30/corr6430.doc