2012

UNED

DISEÑOS DE INVESTIGACIÓN Y ANÁLISIS DE DATOS

[TEMA 8]

Análisis de Regresión Lineal Simple y Múltiple

ÍNDICE

8.1.- Introducción

Como se explica en el libro de Fundamentos de Investigación, "*los diseños **ex post facto** se caracterizan porque el investigador no puede manipular intencionalmente la variable independiente, ni asignar aleatoriamente a los participantes a los diferentes niveles de la misma … en estos diseños, el investigador **selecciona** a los sujetos en función de que posean o no determinadas características*". Uno de los procedimientos de análisis más empleados para este tipo de diseños es el que se conoce como Análisis de Regresión/Correlación. Este procedimiento analítico puede ser usado siempre que una variable cuantitativa, en este caso la Variable Dependiente (VD), sea estudiada como una función de una variable, o de una combinación de varias Variables Independientes[1] (VI). Cuando se estudia la VD en función de una sola VI este análisis se conoce como Análisis de Regresión Simple (ARS). Cuando hay más de una VI se conoce como Análisis de Regresión Múltiple (ARM).

La forma de la relación entre la VD y la VI puede ser muy diversa. En el caso del ARS se pueden dar relaciones lineales, exponenciales, potenciales, polinómicas, etc. En este texto únicamente vamos a tratar las relaciones de carácter lineal, es decir, aquellas en las que la VD se puede expresar como una función de la VI elevada a la primera potencia. Lo mismo sucede con las relaciones que se pueden dar en el ARM, pero sólo estudiaremos el caso en el que la VD se puede expresar como una combinación lineal de varias VI`s.

Aunque el ARM es una técnica de análisis idónea para los diseños **ex post facto**, también se puede aplicar a situaciones en las que se manipulan condiciones experimentales. Por tanto, las variables independientes pueden tener una ocurrencia natural (sexo, Cociente Intelectual, tiempo que se tarda en aprender una lista de palabras, introversión, ansiedad, etc.), o pueden ser variables manipuladas en un laboratorio. En resumen, "*casi cualquier información que tenga interés para el estudio de la VD puede ser objeto de incorporación en este tipo de análisis*"[2].

El Análisis de Regresión tiene una amplitud de aplicación de gran alcance. Se emplea para contrastar hipótesis generadas en el ámbito de las ciencias de la conducta, de la salud, de la educación, etc. Estas hipótesis pueden llegar por la vía de una teoría formal, por investigaciones previas o simplemente por algún tipo de intuición científica acerca de algún fenómeno. Una lista breve de hipótesis sobre determinadas situaciones puede dar idea del alcance de esta técnica de análisis:

- El estrés en la vida cotidiana puede estar relacionado con la cantidad de días que las personas causan baja laboral por enfermedad.

[1] Al igual que en los capítulos de Diseños de más de dos grupos, en este capítulo designaremos la variable dependiente por Y, mientras que las variables independientes las designaremos como X_i, siendo i = 1,….,n, según el número de variables independientes que se incorporen en el ARM.

[2] Cohen, J, Cohen, P. , West, S. G.y Aiken, L. S. *Applied Multiple Regression/Correlation. Analysis for the Behavorial Sciences*. 3ª Ed. Lawrence Erlbaum Assoc. N, Jersey, 2003.

- Cuando, para una política educativa racional, se quiere compara el rendimiento educativo en función de si los estudiantes estudian en colegios públicos o privados, es necesario el control estadístico de determinadas características, tales como el CI, logros académicos previos, formación académica de los padres, nivel de ingresos familiares, etc., porque pueden explicar el rendimiento más que el tipo de escuela.
- La ejecución de una tarea está relacionado con el nivel de activación de las personas, y la relación tiene una forma de U invertida (esta relación se conoce en el ámbito de la psicología experimental como la "**Ley de Yerkes y Dodson**")

Cada una de estas hipótesis plantea una relación entre una o más variables explicativas (VI´s) y la variable dependiente (VD) objeto de estudio y, por consiguiente, todas ellas pueden ser contrastadas mediante Análisis de Regresión.

En este capítulo vamos a estudiar únicamente el Análisis de Regresión Lineal Simple y Múltiple y vamos a apoyar la explicación mediante ejemplos numéricos para facilitar la comprensión de la técnica de análisis, utilizando el mínimo soporte matemático que es posible.

8.2.- Objetivos

- Elaborar un **modelo** de regresión simple, para explicar el comportamiento de una variable (dependiente) a partir de otra (independiente).
- Interpretar los **coeficientes** del modelo elaborado.
- Determinar si el modelo es suficientemente explicativo (bondad de ajuste)
- Especificar el modelo estadístico que subyace al análisis.
- Elaborar un modelo de regresión lineal múltiple con dos variables predictoras.
- Calcular la bondad del modelo de regresión múltiple.
- Realizar inferencias sobre los coeficientes de correlación y los de los modelos de regresión ajustados.
- Cuantificar la correlación de dos variables cuando se excluye el influjo que otras variables tienen sobre cada una de ellas.

8.3.- Análisis de Regresión Simple

Cuando una variable, que llamaremos independiente (VI), aporta información sobre otra variable, que llamaremos dependiente (VD), decimos que ambas están relacionadas y esa información puede servir para saber más sobre el comportamiento de la variable dependiente, sabiendo el comportamiento de la independiente. Esta relación, como se ha señalado en la introducción, puede ser de diversos tipos: lineal, potencial, exponencial, logarítmica, polinómica, etc. El tipo de relación entre las variables se detecta a través de la representación gráfica de todos los pares de valores en ambas variables. Supongamos, por ejemplo, los datos de la Tabla 8.1 (que servirán como conjunto de datos para la explicación del ARS) con las puntuaciones de 16 escolares en dos variables: una prueba de

vocabulario (variable X o independiente) y el número de errores ortográficos detectados dentro de un texto (variable Y o dependiente).

Tabla 8.1

Datos de 16 escolares en una prueba de vocabulario (X) y número de errores ortográficos detectados en un texto (Y)

Sujeto	X	Y	Sujeto	X	Y
1	3	9	9	10	22
2	1	7	10	2	6
3	7	12	11	5	10
4	9	18	12	7	18
5	10	18	13	9	16
6	8	13	14	6	13
7	4	8	15	7	15
8	6	17	16	8	16

Al confeccionar el correspondiente diagrama de dispersión o diagrama de puntos de los 16 pares de datos (véase la Figura 8.1) se observa que hay un tendencia de carácter lineal y positiva, en el sentido que a medida que un escolar puntúa más alto en la prueba de vocabulario (X) también suele detectar más errores ortográficos (Y). Obviamente estamos hablando de una tendencia porque esa relación no siempre se cumple de tal forma que no siempre una mayor puntuación en vocabulario se corresponde con una mayor detección de errores. Véase, por ejemplo, los sujetos 12 y 13; el segundo obtiene una puntuación mayor en la prueba de vocabulario (2 puntos), pero detecta dos errores menos que el primero. Aún así, la **tendencia** global de los datos es claramente directa o positiva.

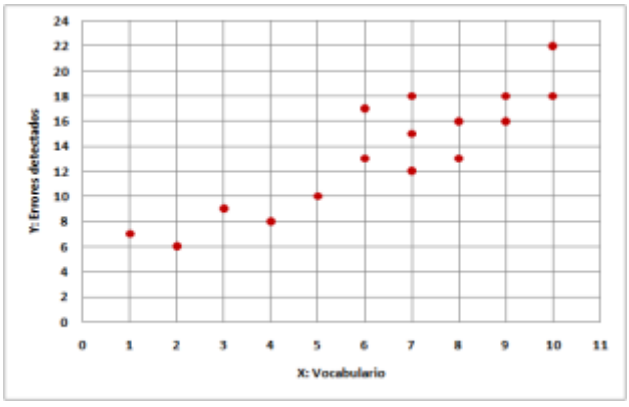

Figura 8.1 *Diagrama de dispersión de los datos de la tabla 8.1*

Por lo estudiado en el texto de Introducción al Análisis de Datos sabemos cómo cuantificar la relación entre dos variables cuantitativas: mediante el Coeficiente de Correlación de Pearson, que puede expresarse en términos de puntuaciones directas, diferenciales o típicas, según las siguientes fórmulas:

$$r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \quad (8.1)$$

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \quad (8.2)$$

$$r_{z_x z_y} = \frac{\sum z_x z_y}{n} \quad (8.3)$$

El resultado del coeficiente con puntuaciones directas y diferenciales para nuestros datos es:

$$r_{xy} = \frac{(16)(1561) - (102)(218)}{\sqrt{[(16)(764) - 102^2][(16)(3294) - 218^2]}} = 0{,}8924$$

$$r_{xy} = \frac{171{,}25}{\sqrt{(113{,}75)(323{,}75)}} = 0{,}8924$$

En la Tabla 8.2 se muestran los cálculos necesarios para obtener los diferentes elementos de las fórmulas. Generalmente la manera más cómoda de calcular el coeficiente de correlación de Pearson será utilizando la fórmula para puntuaciones directas.

Tabla 8.2

Desarrollo para el cálculo del coeficiente de correlación de Pearson

	Puntuaciones directas					Puntuaciones diferenciales				
Sujetos	**X**	**Y**	**XY**	**X^2**	**Y^2**	**x**	**y**	**xy**	**x^2**	**y^2**
1	3	9	27	9	81	-3,375	-4,625	15,609375	11,390625	21,390625
2	1	7	7	1	49	-5,375	-6,625	35,609375	28,890625	43,890625
3	7	12	84	49	144	0,625	-1,625	-1,015625	0,390625	2,640625
4	9	18	162	81	324	2,625	4,375	11,484375	6,890625	19,140625
5	10	18	180	100	324	3,625	4,375	15,859375	13,140625	19,140625
6	8	13	104	64	169	1,625	-0,625	-1,015625	2,640625	0,390625
7	4	8	32	16	64	-2,375	-5,625	13,359375	5,640625	31,640625
8	6	17	102	36	289	-0,375	3,375	-1,265625	0,140625	11,390625
9	10	22	220	100	484	3,625	8,375	30,359375	13,140625	70,140625
10	2	6	12	4	36	-4,375	-7,625	33,359375	19,140625	58,140625
11	5	10	50	25	100	-1,375	-3,625	4,984375	1,890625	13,140625
12	7	18	126	49	324	0,625	4,375	2,734375	0,390625	19,140625
13	9	16	144	81	256	2,625	2,375	6,234375	6,890625	5,640625
14	6	13	78	36	169	-0,375	-0,625	0,234375	0,140625	0,390625
15	7	15	105	49	225	0,625	1,375	0,859375	0,390625	1,890625
16	8	16	128	64	256	1,625	2,375	3,859375	2,640625	5,640625
Suma	**102**	**218**	**1561**	**764**	**3294**	0	0	**171,25**	**113,75**	**323,75**
Media	**6,375**	**13,625**								
Desv. Típ.	**2,666**	**4,498**								

A la vista de los datos representados en el diagrama de la Figura 8.1, es fácil intuir que la relación entre ambas variables puede ser "modelada" de tal forma que la VD se represente como una función de la VI. En este caso, la función que, *a priori* y visto el diagrama, mejor puede modelar la relación es la lineal, es decir, una función que exprese la VD en términos de los valores de la VI, sometidos a algún tipo de transformación lineal. Dicho de otra forma, una función lineal que permita hacer una estimación de la VD a partir de la VI, es una función del tipo:

$$Y' = BX + B_0; expresada\ en\ puntuaciones\ directas \quad (8.4\ a)$$

$$y' = Bx\ \ ; expresada\ en\ puntuaciones\ diferenciales \quad (8.4\ b)$$

$$z'_y = r_{xy}z_x\ \ ; expresada\ en\ puntuaciones\ típicas \quad (8.4\ c)$$

Al ser una estimación, Y' (puntuación en Y predicha por el modelo lineal) se acercará más o menos al verdadero valor de la VD. Este ajuste será mayor cuanto mayor sea la relación entre las variables, es decir, dependerá del valor del coeficiente de correlación de Pearson, como tendremos ocasión de demostrar más adelante. Aún sabiendo que la mejor relación puede ser representada por una función lineal, queda aún por determinar cuál de las muchas funciones lineales (una para cada combinación de valores, parámetros o coeficientes de la regresión, B y B_0 en la Ecuación 8.4a lo cual significa que, en esencia, son infinitas), es la que mejor ajusta los datos del diagrama.

8.3.1 Coeficientes de la regresión lineal simple

Antes de proceder al cálculo de los coeficientes de regresión (B y B_0) es conveniente observar qué sucede una vez que hemos determinado la función y la representamos sobre los datos. En la Figura 8.2 se pueden ver los datos y una línea vertical entre cada uno de los datos y la recta de ajuste que mejor los ajusta (más adelante veremos cómo se calcula esta recta). Cuando ya se ha construido la recta (que es una estimación de Y), y se procede a particularizar para cada valor de la VI (en este caso puntuación en vocabulario), los valores resultantes se sitúa, obviamente, a lo largo de la recta. En algunos casos el valor que se obtiene con la recta de ajuste (la estimación, Y') coincide con el verdadero valor de la VD (representado por los puntos), aunque en la mayoría de los casos no coincide. Es decir, si deseamos predecir el comportamiento de VD utilizando su relación con VI, una vez hecha la predicción (valor en la recta), vemos que en muchos casos difiere del verdadero valor de la VD para ese valor concreto de la VI. Por tanto, cuando utilizamos el modelo lineal para estimar cada valor Y a partir de X aplicando la recta de regresión obtenida, **hay un error en la estimación de la VD (Y)** ya que el valor pronosticado (Y') y el valor medido (Y) no suelen coincidir. La diferencia entre ambos es ese error de estimación. En la Figura 8.2 este error viene dado por la magnitud o longitud de la línea vertical que separa cada dato de la predicción realizada por la recta de regresión.

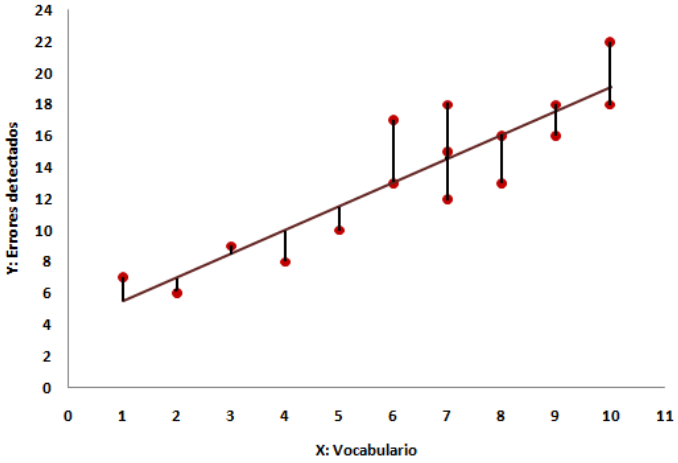

Figura 8.2 *Errores después del ajuste de una recta*

De acuerdo a la fórmula (8.4a), los valores en la recta los hemos denominado Y', y a los valores de la VD los hemos denominado Y. Pensemos en estas distancias (Y – Y'), como la distancia que hay entre cada valor (Y) y su media (representada por Y', ya que la predicción realizada por la recta de regresión representa la media que sería de esperar si el análisis se repitiese con infinitas muestras). Ahora, tomemos estas distancias, elevémoslas al cuadrado y sumemos todos esos cuadrados. El valor resultante de esta suma será el Error Cuadrático de la Recta de Ajuste (existen otras terminologías como Recta de Estimación, Recta de Predicción o Recta de Regresión, siendo cualquiera de estas denominaciones es válida), y sólo hay una recta que hace mínimo este error. Por esta razón a este método de ajuste de una recta de regresión se le conoce como **ajuste por mínimos cuadrados** ya que el objetivo es encontrar los valores B y B_0 que hacen más pequeño (mínimo) el error (Y-Y') al cuadrado.

Además, hay otra característica importante de la recta de ajuste, que se puede enunciar del siguiente modo: la recta de regresión es una estimación insesgada de la VD en el sentido de que la media de los valores pronosticados es igual a la media de los valores observados. Es decir,

$$\frac{\sum Y_i}{n} = \frac{\sum Y'_i}{n} \rightarrow \bar{Y} = \bar{Y}' \tag{8.5}$$

Por procedimientos matemáticos que no vamos a desarrollar, el valor del parámetro *B* de la función lineal en (8.4) que minimiza los errores cuadráticos, se obtiene de acuerdo a la expresión:

$$B = r_{XY}\frac{S_Y}{S_X} \tag{8.6}$$

Siendo:

r_{XY}, el coeficiente de correlación de Pearson

S_Y la desviación típica de la variable dependiente (Y)

S_X la desviación típica de la variable independiente (X).

Conocido *B*, el valor de B_0 se obtiene mediante la expresión:

$$B_0 = \bar{Y} - B\bar{X} \tag{8.7}$$

Construida la recta de ajuste podemos expresar la variable dependiente, *Y*, como una función de la variable independiente, *X*, mediante la siguiente expresión:

$$Y = B_0 + BX + \varepsilon \tag{8.8}$$

Donde ε representa el error de predicción y está compuesto por las distancias entre cada valor de *Y* e *Y'* para una valor dado de *X* que observaríamos si repitiésemos el procedimiento a varias muestras diferentes.

¿Cuál es el significado de los coeficientes de regresión? En el análisis de regresión simple el coeficiente "protagonista" es el factor *B*, conocido como **pendiente de la recta**, y cuantifica el incremento que se produce en la estimación de la variable dependiente (*Y'*) cuando la independiente (*X*) aumenta en una unidad.

En la Figura 8.3 se ve de manera gráfica el significado de B en nuestros datos. La estimación de Y para un valor X = 4, proporciona el valor 10,049, y para una X = 5, el valor es 11,555. La diferencia entre estos valores al aumentar X en una unidad (de 4 a 5) es lo que aumenta Y' y ese es el valor de la pendiente. En el caso del ejemplo que ilustra esta explicación la pendiente nos dice que los escolares, con cada punto más que obtienen en la prueba de vocabulario detectan, en promedio, 1,5 errores más en la prueba de lectura.

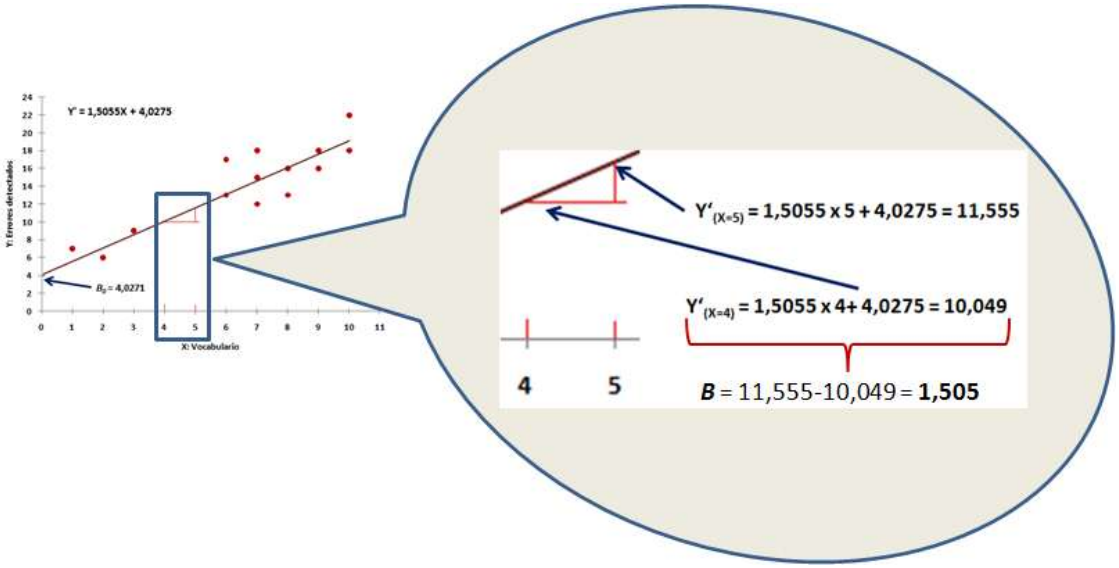

Figura 8.3 *Interpretación gráfica de la pendiente de la recta de regresión*

La constante de la recta de regresión, B_0, señala el punto en el que ésta corta al eje de ordenadas, es decir, el valor estimado de Y cuando X es igual a 0. No es un coeficiente interpretable en el sentido en el que lo es la pendiente. De hecho, casi nunca es objeto de interpretación salvo cuando el valor 0 se encuentra dentro del rango de valores de la VI. Si no es el caso, la recta de regresión sólo se puede interpretar dentro del rango de valores de la VI, pues es con esos valores con los que se construye la recta de estimación. Fuera de ese rango, no se sabe qué sucede con la función que relaciona X con Y y por tanto podría ser que por debajo del menor valor de la VI y/o por encima del mayor valor de la VI la función de estimación de la VD cambiara su forma.

Para que sean válidas las inferencias que sobre la VD se hagan con la recta de regresión, se deben de cumplir cuatro **supuestos básicos**, tres de los cuales son, en esencia, los mismos que ya se han mencionado en las técnicas de análisis para las pruebas T y los ANOVAS:

1. Independencia de las observaciones. Este supuesto sólo se contrasta si el proceso de selección de la muestra no ha sido aleatorio.
2. Homocedasticidad. Su cumplimiento supone que las varianzas de las distribuciones de los errores, condicionadas a los diferentes valores de la VI, deben ser iguales.
3. Normalidad de las distribuciones condicionadas.
4. Independencia entre los valores estimados, Y', y los errores de estimación, ε. Expresado en términos de coeficiente de correlación de Pearson, $\mathbf{r_{Y'\varepsilon} = 0}$. Esto es así debido a que los errores se distribuyen de manera aleatoria, mientras que las estimaciones o pronósticos son una función de la VI.

En la Figura 8.4 se representan los supuestos 2 (las varianzas de las cuatro curvas normales dibujadas son idénticas) y 3 (para cada valor de X_i existe una gama de valores posibles que se distribuyen normalmente con media Y').

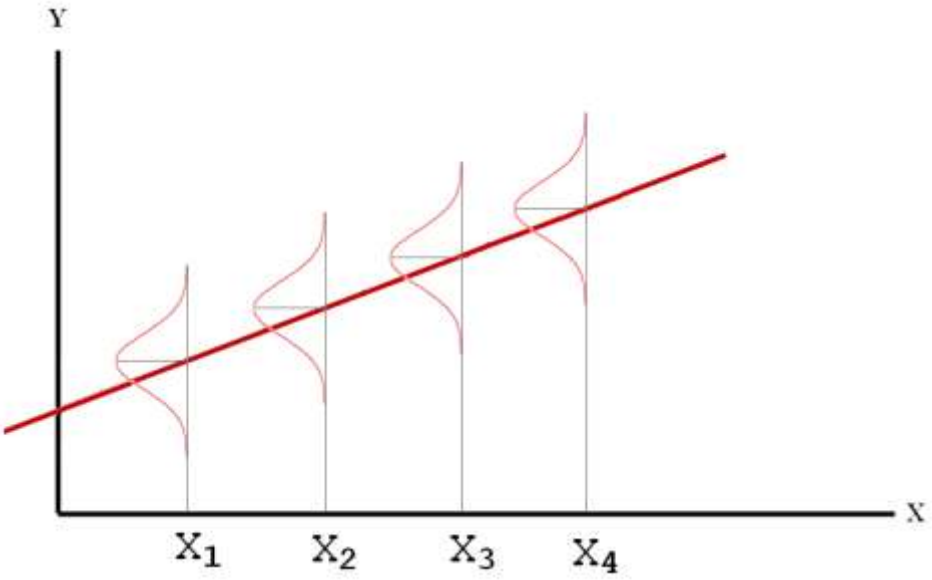

Figura 8.4 *Representación supuestos 2 y 3 en el ARS*

8.3.2 Bondad de Ajuste de la Recta de Regresión

La expresión Bondad de Ajuste, se refiere a cómo de "explicativa" es la recta respecto de los datos sobre los que se ha ajustado. Al hacer un ajuste mínimo cuadrático conseguimos un conjunto de valores, situados sobre la recta, cuyo promedio coincide con el promedio de la VD, que estiman los diferentes valores de la VD para cada valor de la VI. Denotaremos simbólicamente estos valores estimados mediante el símbolo de la variable dependiente (v.g., Y) con un acento en la parte superior derecha, es decir, como Y' y la nombraremos diciendo "Y prima". Las estimaciones pueden diferir de los valores de la VD, es decir, los valores de Y no tienen porqué coincidir exactamente con Y'. La diferencia entre ambos valores será un "error de estimación" que, siendo inevitable, trataremos de que sea lo menor posible. La magnitud de los errores de estimación son un primer indicio para determinar si el ajuste es bueno o no. No obstante, tomar la magnitud de los errores aisladamente, sin poner ésta en relación con alguna otra magnitud, no resuelve completamente el problema de determinar la bondad.

Para explicar el concepto de bondad de ajuste, veamos de qué está compuesta la varianza de la VD, antes y después de ajustar la recta de regresión sobre el conjunto de datos. Para ello, vamos a estudiar lo que sucede en uno solo de los 16 valores que estamos utilizando como ejemplo numérico, tal como se observa en la Figura 8.5.

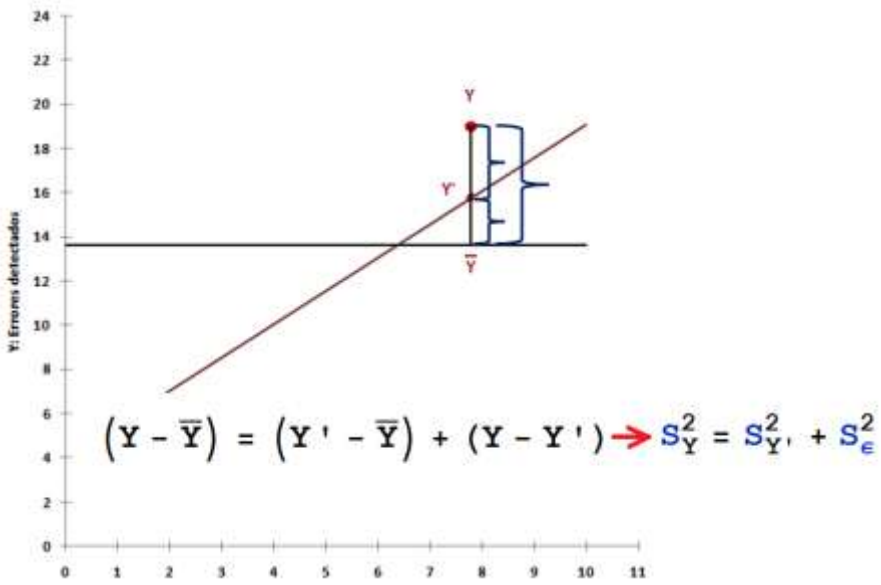

Figura 8.5 *Descomposición de la suma de cuadrados de la VD*

Imagine el lector que sólo dispone de los estadísticos media y varianza de la prueba de detección de errores del grupo de sujetos del ejemplo de la Tabla 8.1, y desea hacer una estimación para un sujeto concreto, y aún no sabe que se ha determinado que dicha prueba está relacionada con la prueba de vocabulario. A falta de otro tipo de información sobre la variable que se quiere estimar, lo correcto es otorgar como mejor estimación la media del grupo, por ser la que minimiza el error de predicción. Para cada sujeto concreto el error que se cometerá será $(Y_i - \bar{Y})$, y para el conjunto de datos, el error, expresado en términos de cuadrados, es el mínimo posible. Es decir, si no tenemos idea de la relación entre la variable predictora (X) y la predicha (Y), y nos piden realizar una estimación de Y, lo mejor que podemos hacer es utilizar la media de Y como estimador. Cualquier otro valor que elijamos hará que el error de predicción sea superior al que obtendríamos si utilizamos la media.

Sigamos suponiendo que en un momento posterior disponemos de información sobre la relación que hay entre la detección de errores (Y) y una prueba de vocabulario (X) que han realizado los estudiantes, y sabe cuál es la recta de ajuste entre ambas variables. Si ahora desea hacer una estimación de la puntuación de un sujeto en la prueba de errores, lo razonable es que aproveche la información nueva de que dispone, y vea qué puntuación ha obtenido el sujeto en vocabulario y, mediante la ecuación de la recta construida, haga una estimación de la puntuación en Y, que estará situada en la recta y que hemos denominado Y'. Esta estimación se aproxima más al valor original (Y) de lo que lo hacía la media pero aún persiste un cierto error, el que va de Y a Y'. Es decir, del error original, $(Y - \bar{Y})$, hemos reducido una parte, $(Y' - \bar{Y})$, pero aún queda otra parte, $(Y - Y')$, sin explicar. Por tanto, la variable original *Y*, expresada en puntuaciones diferenciales, $(Y - \bar{Y})$, es la suma de otras dos variables, $(Y - Y')$ y $(Y' - \bar{Y})$, que, además, son independientes entre sí; su correlación, pues, es cero. Por tanto podemos afirmar que a nivel de cada puntuación individual se cumple:

$$(Y - \bar{Y}) = (Y' - \bar{Y}) + (Y - Y') \tag{8.9}$$

y sumando para todos los puntos y elevando al cuadrado se obtiene lo que se conoce como Suma de Cuadrados, dividiendo por el número de casos se obtienen la varianza total de Y (S_Y^2), la varianza de las Y predichas ($S_{Y'}^2$) y la varianza de los errores (S_ε^2). Como la relación de la Ecuación 8.9 se sigue manteniendo, estas varianzas mantienen la relación que puede verse en la Ecuación 8.10:

$$\frac{\sum(Y - \bar{Y})^2}{n} = \frac{\sum(Y' - \bar{Y})^2}{n} + \frac{\sum(Y - Y')^2}{n} \rightarrow S_Y^2 = S_{Y'}^2 + S_\varepsilon^2 \tag{8.10}$$

En resumen, cuando hay una relación lineal entre dos variables, la varianza de la VD se puede descomponer en dos varianzas: la de los pronósticos, debido a la relación que la VD guarda con la VI, y la de los errores o residuos. Esta relación se cumple tanto para la Regresión Lineal Simple como para la Múltiple. Esta descomposición de las varianza de la VD en dos varianzas es el "Teorema de Pitágoras" del Análisis de Regresión Lineal.

Tabla 8.3

Desarrollo numérico de la descomposición de la varianza de la VD

X	Y	$\bar{Y}$	Y'	$(Y - \bar{Y})$	$(Y' - \bar{Y})$	$(Y - Y')$	$(Y - \bar{Y})^2$	$(Y' - \bar{Y})^2$	$(Y - Y')^2$
3	9	13,6250	8,5440	-4,6250	-5,0810	0,4560	21,3906	25,8170	0,2080
1	7	13,6250	5,5330	-6,6250	-8,0920	1,4670	43,8906	65,4810	2,1522
7	12	13,6250	14,5659	-1,6250	0,9409	-2,5659	2,6406	0,8854	6,5840
9	18	13,6250	17,5769	4,3750	3,9519	0,4231	19,1406	15,6177	0,1790
10	18	13,6250	19,0824	4,3750	5,4574	-1,0824	19,1406	29,7834	1,1716
8	13	13,6250	16,0714	-0,6250	2,4464	-3,0714	0,3906	5,9850	9,4337
4	8	13,6250	10,0495	-5,6250	-3,5755	-2,0495	31,6406	12,7846	4,2002
6	17	13,6250	13,0604	3,3750	-0,5646	3,9396	11,3906	0,3187	15,5201
10	22	13,6250	19,0824	8,3750	5,4574	2,9176	70,1406	29,7834	8,5123
2	6	13,6250	7,0385	-7,6250	-6,5865	-1,0385	58,1406	43,3825	1,0784
5	10	13,6250	11,5549	-3,6250	-2,0701	-1,5549	13,1406	4,2851	2,4179
7	18	13,6250	14,5659	4,3750	0,9409	3,4341	19,1406	0,8854	11,7928
9	16	13,6250	17,5769	2,3750	3,9519	-1,5769	5,6406	15,6177	2,4867
6	13	13,6250	13,0604	-0,6250	-0,5646	-0,0604	0,3906	0,3187	0,0037
7	15	13,6250	14,5659	1,3750	0,9409	0,4341	1,8906	0,8854	0,1884
8	16	13,6250	16,0714	2,3750	2,4464	-0,0714	5,6406	5,9850	0,0051
						Suma	323,7500	257,8159	65,9341
						Varianzas	**20,234**	**16,1135**	**4,1209**

A partir de la Ecuación 8.10, se puede establecer una serie de relaciones. La primera es lo que representa la proporción de la varianza de los pronósticos respecto de la VD: ***la proporción de la varianza de la VD explicada por la varianza de la VI***, ya que los pronósticos son un combinación lineal de la propia VI, combinación que está representada por la recta de regresión (***Y' = BX + B_0***). La cuantía de esta proporción es el cuadrado del coeficiente de correlación de Pearson entre la VD y la VI (esto solo sirve para el caso de la Regresión Lineal Simple).

$$\frac{S_{Y'}^2}{S_Y^2} = \frac{\sum(Y' - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{SC_{reg}}{SC_Y} = r_{XY}^2 \quad (8.11)$$

$$\frac{S_\varepsilon^2}{S_Y^2} = \frac{\sum(Y - Y')^2}{\sum(Y - \bar{Y})^2} = \frac{SC_{residuos}}{SC_Y} = 1 - r_{XY}^2 \quad (8.12)$$

En resumen, r_{XY}^2 (que también designaremos como R^2), denominado ***Coeficiente de Determinación***, es la proporción de la variabilidad de la VD que es imputada (o explicada por) la variabilidad de la VI, mientras que su complemento, $(1 - r_{XY}^2)$, denominado ***Coeficiente de Alienación***, es la parte residual de la variabilidad de la VD, atribuible a otros factores no relacionados linealmente con la VD.

Además de esta interpretación de R^2, hay otra que tiene que ver con la reducción del error original de la VD. En este sentido, ***R^2 es la proporción en que se reduce el error de la VD cuando empleamos la recta de regresión para estimarla***. Observe el lector (Tabla 8.3) que el error cuadrático inicial es 323,75, y después de ajustar la recta y proceder a las estimaciones de Y, aún queda un error cuadrático de 65,9341. En términos absolutos el error se ha reducido en 323,75 – 65,9341 = 257,8159, lo que en términos de proporción respecto del error original la reducción es: 257,8159/323,75 = 0,7963, que es el valor de R^2 que aparece en la Tabla.

A partir de 8.12, se puede obtener la desviación estándar de los errores (o residuos). Su expresión es:

$$\frac{S_\epsilon^2}{S_Y^2} = 1 - r_{xy}^2 \rightarrow S_\varepsilon = S_Y = \sqrt{1 - r_{XY}^2} \quad (8.13)$$

Un forma gráfica de representar la varianza explicada o compartida es mediante los denominados diagramas de Venn en estadística matemática, en el cual la varianza de cada variable es representada por sendos círculos de área igual a la unidad y la intersección del solapamiento de ambos círculos representaría la proporción de varianza compartida, que es el valor del coeficiente de determinación R^2. En la Figura 8.6 se representa la varianza compartida de los datos del ejemplo, sin pretensión de exactitud en cuanto al área solapada de ambos círculos.

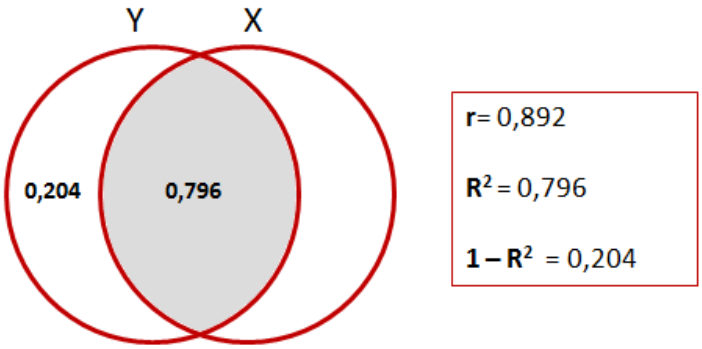

Figura 8.6 *Diagrama de Venn con la representación de la proporción de varianza compartida*

Otro indicador del ajuste, además de R^2, es lo que se conoce como Error Típico, y es una estimación sobre la población realizada a partir de la muestra. Su valor se deriva de la raíz cuadrado del cociente entre la Suma de Cuadrados de los residuos o errores entre los grados de libertad, que son el número de observaciones muestrales menos el número de parámetros estimados por la regresión, que en el caso bivariado son dos. La fórmula es:

$$\hat{S}_{\varepsilon} = \sqrt{\frac{\sum(Y - Y')^2}{n - p - 1}} \tag{8.14}$$

Siendo p, el número de variables independientes que incorpora el modelo, que en el caso de la regresión lineal simple es 1.

8.3.3.- Inferencias sobre correlación y regresión

Una vez construido el modelo de estimación, es preciso dotarle de significación estadística para que las inferencias que se hagan a partir de los datos muestrales sean válidas para el conjunto de la población. Los dos contrastes que vamos a tratar son los que tienen que ver con el coeficiente de correlación entre las variables dependiente e independiente, y por tanto también es un contraste sobre la regresión, y el segundo es el contraste que se realiza sobre los coeficientes de regresión. Además del contraste, veremos cómo calcular los intervalos de confianza tanto para el coeficiente de correlación como para los coeficientes de la regresión.

8.3.3.1.- Contraste sobre el coeficiente de correlación de Pearson.

El contraste de hipótesis que se presenta a continuación, es relativo al ajuste de la correlación entre la VD y la VI. En este caso la hipótesis nula será que no hay relación lineal entre la VD y la VI, siendo la hipótesis alternativa su negación, es decir, que sí hay relación.

Condiciones y supuestos. Tenemos dos variables medidas en una escala de intervalo o razón que se distribuyen normalmente en la población. En el caso del ejemplo de este capítulo, hemos de suponer que las variables prueba de vocabulario (X) y el número de errores ortográficos (Y) se distribuyen normalmente en la población.

Formulación de hipótesis. La hipótesis nula ha de postular que en la población el coeficiente de correlación de Pearson es igual a cero, mientras que la hipótesis alternativa indica que la relación lineal entre X e Y es significativa:

$$H_0\colon\ \rho = 0$$
$$H_1\colon\ \rho \neq 0$$

Estadístico de contraste y distribución muestral. El estadístico de contraste se distribuye según t de Student con $n-2$ grados de libertad, y viene dado por la siguiente fórmula:

$$t = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}} \tag{8.15}$$

Con los datos del ejemplo del presente tema:

$$t = \frac{0{,}8924\sqrt{16-2}}{\sqrt{1-0{,}8924^2}} = 7{,}3988 \approx 7{,}40$$

Establecer regla de decisión en función del nivel de confianza. Para un nivel de confianza del 99% en un contraste bilateral, el valor crítico obtenido en las tablas t de Student es igual a: 2,977. Dado que: $7{,}40 > 2{,}977$ rechazamos la hipótesis nula, concluyendo que la relación entre X e Y es significativa.

Mediante un programa informático adecuado se comprueba que el nivel crítico es: $p < 0{,}0001$. Con las tablas se llega a la conclusión que el nivel crítico es: $p < 0{,}005$, que es la probabilidad de obtener valores superiores a 2,977 en una distribución t de Student con 14 grados de libertad.

Interpretar los resultados en función del contexto de la investigación. Existe relación lineal entre las variables prueba de vocabulario (X) y número de errores ortográficos (Y).

Podemos observar que si el coeficiente de correlación de Pearson es distinto de cero, también será distinta de cero la pendiente de la recta de regresión de Y sobre X, dado que ambos índices se relacionan según la siguiente ecuación:

$$B = r_{XY}\frac{S_Y}{S_X}$$

8.3.3.2.- Contraste para el coeficiente de regresión B (ANOVA).

En el caso de la regresión lineal simple, se puede contrastar si la pendiente de la recta de regresión es significativa, utilizando la descomposición de la variabilidad total vista en el apartado 8.3.2. Ordenando los datos en una tabla como las utilizadas en los temas sobre análisis de varianza tenemos:

Tabla 8.4

Tabla ANOVA para el contraste de la regresión

Fuentes de variación	Sumas de cuadrados	Grados de libertad	Medias cuadráticas	F
Regresión	SC_{Reg}	1	$MC_{Reg} = \frac{SC_{Reg}}{1}$	$F = \frac{MC_{Reg}}{MC_{Res}}$
Residual	SC_{Res}	$n-2$	$MC_{Res} = \frac{SC_{Res}}{n-2}$	
Total	SC_{Total}	$n-1$		

Podemos completar la Tabla 8.4 con las sumas de cuadrados calculadas en la Tabla 8.3.

FV	SC	gl	MC	F
Regresión	257,816	1	257,816	54,7
Residual	65,934	14	4,709	
Total	323,750	15		

También podríamos calcular el estadístico F mediante la siguiente expresión:

$$F = \frac{R^2/1}{(1-R^2)/(N-2)} = \frac{0{,}7963}{(1-0{,}7963)/(16-2)} = 54{,}743 \qquad (8.16)$$

El estadístico de contraste F resulta significativo, pues la probabilidad de encontrar un valor F igual o mayor, con 1 y 14 grados de libertad es p = 3,358x10^{-6} (valor calculado mediante un programa informático).

El estadístico de contraste T visto en el punto 8.3.3.1 y el estadístico F están relacionados según la siguiente expresión:

$$t_n^2 = F_{1,n} \qquad (8.17)$$

8.3.3.3.- Contraste para el coeficiente de regresión B.

Otra forma de determinar si hay evidencia estadística de que la pendiente es diferente de cero, es decir si la pendiente es significativamente diferente a una línea horizontal, perpendicular al eje de ordenadas es la siguiente:

Condiciones y supuestos. Tal y como vimos en el punto 8.3.1, los supuestos son: Independencia de las observaciones, homocedasticidad, normalidad de las distribuciones condicionadas e independencia entre los valores estimados y los errores de estimación.

Formulación de hipótesis. Normalmente estaremos interesados en comprobar si la pendiente de la recta de regresión en la población es distinta de cero.

$$H_0: \quad \beta = 0$$
$$H_1: \quad \beta \neq 0$$

Estadístico de contraste y distribución muestral. El estadístico de contraste se distribuye según t de Student con $n-2$ grados de libertad, y viene dado por la siguiente expresión:

$$t = \frac{B-\beta}{\sigma_B} = \frac{B-\beta}{\frac{S_Y}{S_X}\sqrt{\frac{1-r_{XY}^2}{(n-2)}}} \tag{8.18}$$

Siendo β el valor especificado en la hipótesis nula. Normalmente, como en este caso, estaremos interesados en comprobar si: $\beta = 0$. Aplicando este contraste a la pendiente de los datos que están sirviendo de ejemplo, el valor del estadístico es:

$$t = \frac{1{,}5055-0}{\frac{4{,}498}{2{,}666}\sqrt{\frac{1-0{,}8924^2}{(16-2)}}} = \frac{1{,}5055}{0{,}2035} \approx 7{,}4$$

Observe el lector que el valor obtenido en este caso es igual al estadístico t utilizado en el punto 8.3.3.1. Efectivamente, siempre que $\beta = 0$:

$$t = \frac{B-0}{\frac{S_Y}{S_X}\sqrt{\frac{1-r_{XY}^2}{(n-2)}}} = \frac{r_{XY}\frac{S_Y}{S_X}}{\frac{S_Y}{S_X}\sqrt{\frac{1-r_{XY}^2}{(n-2)}}} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

Establecer regla de decisión en función del nivel de confianza. Para un nivel de confianza del 99% en un contraste bilateral: $7{,}40 > 2{,}977$, luego rechazamos la hipótesis nula, concluyendo que la pendiente de la ecuación de regresión es distinta de cero, siendo el nivel crítico: $p < 0{,}0001$

Interpretar el resultado en función del contexto de investigación. Existe relación lineal entre la prueba de vocabulario (X) y el número de errores ortográficos detectados en un texto (Y), de manera que podemos pronosticar los valores de la VD en función de los valores de la VD.

Intervalo de confianza. El intervalo de confianza para la pendiente de la recta de regresión se puede calcular mediante la siguiente expresión:

$$IC(B) = B \pm \left(t_{(n-2;1-\alpha/2)}\right)(\sigma_B) \tag{8.19}$$

Aplicando la fórmula a los resultados del ejemplo se obtiene, para un nivel de confianza del 95%, los siguientes límites:

$$IC_B = 1{,}5055 \pm (2{,}145)\left(\frac{4{,}498}{2{,}666}\sqrt{\frac{1-0{,}8924^2}{(16-2)}}\right) = \begin{cases}1{,}942\\1{,}069\end{cases}$$

8.3.3.4.- Contraste para el coeficiente de regresión B_0.

También se puede comprobar si el intercepto es distinto de cero, aunque en este caso, ya se ha señalado que en la mayor parte de los estudios suele ser ignorado. El estadístico de contraste se distribuye según t de Student con $n-2$ grados de libertad, y viene dado por la expresión:

$$t = \frac{B_0 - 0}{\sigma_{B_0}} = \frac{B_0 - 0}{\sigma_\varepsilon\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\hat{S}_X^2}}} \tag{8.20}$$

Siendo σ_ε el ***Error Típico***, ya comentado en el epígrafe de bondad de ajuste, y cuyo valor es la raíz cuadrada de la Media Cuadrática (MC) de los Residuos de la tabla del ANOVA (Tabla 8.4) para el contraste de la regresión, que representa la varianza residual en la población para el caso de la regresión bivariada. Como en el caso de la pendiente, el estadístico *t* tiene la misma distribución con los mismos grados de libertad.

Aplicando el contraste a los datos del ejemplo:

$$t = \frac{4{,}0275 - 0}{\sqrt{4{,}7096}\sqrt{\frac{1}{16} + \frac{6{,}375^2}{(16-1)7{,}583}}} = \frac{4{,}0275}{1{,}4061} = 2{,}864$$

Con $\alpha = 0{,}05$ en un contraste bilateral rechazamos la hipótesis nula de que el intercepto es igual a 0 ya que en este caso, para 14 grados de libertad los valores críticos son: -2,14 y 2,14

Para el intercepto, la fórmula de cálculo del intervalo de confianza es:

$$IC(B_0) = B_0 \pm (t_{(n-2;1-\alpha/2)})(\sigma_{B_0}) \qquad (8.21)$$

Aplicando la expresión a los datos del ejemplo los límites son:

$$IC_{B_0} = 4{,}0275 \pm (2{,}145)\left(\sqrt{4{,}7096}\sqrt{\frac{1}{16} + \frac{6{,}375^2}{(16-1)7{,}583}}\right) = \begin{cases} 7{,}043 \\ 1{,}012 \end{cases}$$

8.4.- Análisis de Regresión Múltiple

Como se ha señalado en el epígrafe de Introducción, en este tema sólo tratamos modelos lineales de explicación del comportamiento de una VD en función de una o varias VI. Ya hemos desarrollado la técnica de **Análisis de Regresión Lineal Simple**, y en este epígrafe ampliamos dicho modelo para más de una VI, empezando por dos VI o variables predictoras. Como en el caso de una sola variable predictora, se va a desarrollar con el mínimo aparato matemático posible. La técnica de cálculo con el modelo de dos variables independientes es relativamente sencilla y se puede desarrollar con un calculadora científica, aunque su modelo matemático, el mismo que el del **Modelo Lineal General** (MGL), del cual los modelos de regresión y los modelos de análisis de la varianza son parte, requiere para su desarrollo algebra de matrices, el cual queda fuera del alcance de este texto. Dado que, en la actualidad, todos estos procedimientos de análisis se realizan con programas informáticos de análisis estadístico, el interés estriba en saber leer e interpretar correctamente los resultados del análisis. Comenzaremos, con el modelo más simple de regresión lineal múltiple que es el de dos variables independientes.

8.4.1.- Regresión con dos Variables Independientes

Para la explicación vamos a servirnos de un ejemplo numérico que hace menos abstracto el modelo. Supongamos que un psicólogo escolar quiere determinar qué factores pueden influir en el rendimiento en matemáticas en uno de los cursos de educación secundaria. Supone que el tiempo que dedican al estudio en general es importante, y quizás también su capacidad para el razonamiento abstracto. Para llevar a cabo esta investigación, selecciona al azar una muestra de 15 estudiantes del colegio y registra el tiempo semanal de estudio (variable X_1) y les administra, además, un test de razonamiento abstracto (variable X_2). Las notas obtenidas por estos 15 escolares en el último examen que han realizado de matemáticas le sirven como variable dependiente (Y). Los datos son los que se muestran en la Tabla 8.5

Tabla 8.5

Datos para el desarrollo del análisis con dos VI

Sujeto	Horas Estudio (X_1)	Test Razonamiento (X_2)	Punt. Matemáticas (Y)
1	8	19	54
2	9	18	52
3	6	14	34
4	9	24	63
5	9	19	46
6	9	16	44
7	12	17	50
8	9	14	52
9	6	23	57
10	11	21	53
11	10	17	56
12	13	19	67
13	9	24	57
14	9	19	54
15	11	17	51

El modelo de estimación lineal de la VD con dos VI's, constará de dos coeficientes de regresión, uno para cada VI, y una constante que será el valor estimado para la VD cuando son nulas las dos VI. No obstante, como ya hemos explicado anteriormente, la constante, si no está el valor cero dentro del rango de valores de las variables predictoras no se toma en consideración en el análisis. Es decir, si X_1= 0 y X_2 = 0 no forman parte de los rangos admitidos empíricamente por ambas variables, no tiene sentido considerar el valor que adoptaría la constante en esos casos. El modelo de estimación es:

$$Y' = B_1X_1 + B_2X_2 + B_0 \tag{8.22}$$

Por lo que la VD se puede expresar como:

$$Y = Y' + \varepsilon \; = \; B_1X_1 + B_2X_2 + B_0 + \varepsilon \tag{8.23}$$

Siendo B_1 el coeficiente de regresión parcial para X_1, B_2 el coeficiente de regresión parcial para X_2, y B_0 el intercepto con el eje de la Y cuando X_1 y X_2 valen 0, y ε los residuos una vez que se ha determinado la función de estimación de la VD. Al igual que en regresión simple, estos coeficientes son los que hacen mínimo el error cuadrático de predicción, es decir, minimizan las diferencias cuadráticas entre Y e Y'.

En primer lugar, antes de calcular los coeficientes de regresión parciales de la ecuación, llamados así para remarcar que es el peso o efecto de una VI cuando el resto de las VI que están en la ecuación permanecen constantes, en la Tabla 8.6 se muestran los estadísticos descriptivos de cada una de las

variables, los coeficientes de correlación entre las variables dos a dos (también llamados bivariados) y las rectas de regresión simple entre cada predictor y la VD. Hemos simplificado la notación de los coeficientes de correlación (r_{y1} representa la correlación entre la variable Y y el predictor X_1, y el resto siguen la misma pauta) y también de la regresión ($Y_1^{'}$ representa las estimaciones Y realizadas a partir de X_1)

Tabla 8.6

Estadísticos descriptivos de los datos de la Tabla 8.5

	Horas Estudio (X_1)	Test Razonamiento (X_2)	Punt. Matemáticas (Y)	Rectas de Regresión
Media	9,33	18,73	52,67	
Desv. Típic.	1,850	3,065	7,498	$Y'_1 = 1{,}786\ X_1 + 36$
r_{y1}	0,441	$r^2_{Y1} = 0{,}194$		$Y'_2 = 1{,}537\ X_2 + 23{,}867$
r_{y2}	0,628	$r^2_{Y2} = 0{,}394$		
r_{12}	-0,043	$r^2_{12} = 0{,}002$		

Para facilitar el cálculo de los coeficientes de regresión parcial de la ecuación, comenzaremos, por sencillez, obteniendo la ecuación de regresión en puntuaciones típicas o estandarizadas, cuya expresión es:

$$z'_y = \beta_1 z_1 + \beta_2 z_2 \tag{8.24}$$

siendo β_1 y β_2 los coeficientes de regresión parcial estandarizados, y se obtienen mediante las siguientes fórmulas:

$$\beta_1 = \frac{r_{Y1} - r_{Y2} r_{12}}{1 - r_{12}^2}$$

$$\beta_2 = \frac{r_{Y2} - r_{Y1} r_{12}}{1 - r_{12}^2} \tag{8.25}$$

Con los resultados de la Tabla 8.6

$$\beta_1 = \frac{0{,}441 - (0{,}628)(-0{,}043)}{1 - (-0{,}043)^2} = 0{,}469$$

$$\beta_2 = \frac{0{,}628 - (0{,}441)(-0{,}043)}{1 - (-0{,}043)^2} = 0{,}649$$

Una vez determinados los coeficientes de regresión parcial estandarizados, se obtienen fácilmente los coeficientes sin estandarizar mediante la relación:

$$B_1 = \beta_1 \frac{S_Y}{S_1}$$

$$B_2 = \beta_2 \frac{S_Y}{S_2} \qquad (8.26)$$

siendo S_1 y S_2, las desviaciones típicas de las variables X_1 y X_2, respectivamente. Los coeficientes no estandarizados son:

$$B_1 = 0{,}469 \frac{7{,}498}{1{,}85} = 1{,}899$$

$$B_2 = 0{,}649 \frac{7{,}498}{3{,}065} = 1{,}587$$

Y la constante de la ecuación es:

$$B_0 = \bar{Y} - B_1 \bar{X}_1 - B_2 \bar{X}_2 \qquad (8.27)$$

Sustituyendo por los valores correspondientes su valor es:

$$B_0 = 52{,}67 - (1{,}899)(9{,}33) - 1{,}587(18{,}73) = 5{,}217$$

Obtenidos los coeficientes, las funciones de estimación de la VD con coeficientes de regresión parcial no estandarizados y estandarizados (es decir, expresada la función en puntuaciones directas y típicas), son las siguientes:

$$Y' = 1{,}899X_1 + 1{,}587X_2 + 5{,}217$$

$$z'_Y = 0{,}469z_1 + 0{,}649z_2$$

Al ser dos las variables independientes, las estimaciones quedan situadas en un plano, que se conoce como plano de regresión, del mismo modo que la línea de estimación en regresión simple se conoce como línea de regresión. Algunas de las puntuaciones de la VD estarán por encima del plano y otras por debajo, y esas distancias de cada punto de la VD al plano forman los residuos del modelo de estimación (véase Figura 8.7).

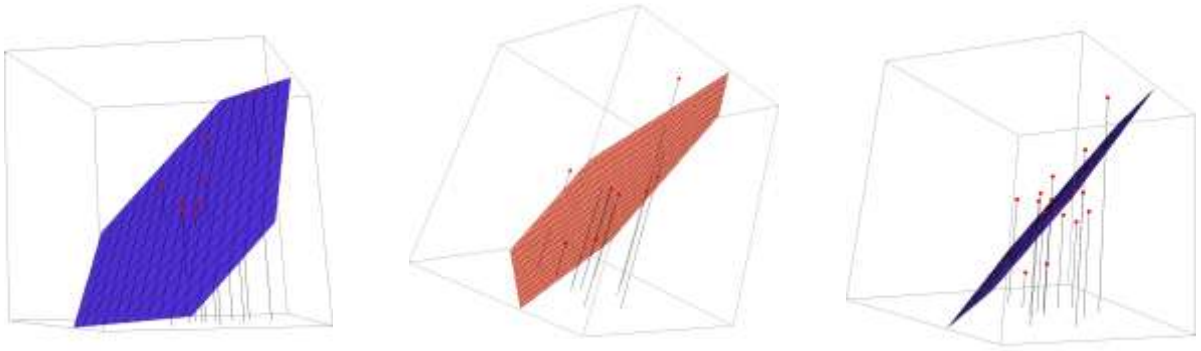

Figura 8.7: *tres vistas del conjunto de puntos y el plano de regresión. La zona azul representa el plano visto "desde arriba", la zona naranja representa el plano visto "desde abajo" . La tercera gráfica intenta visualizar todos los puntos, tanto los que están situados por encima como los que están situados por debajo del plano. En este caso, el plano se ve en "escorzo". Los datos están representados por puntos rojos.*

El modelo ajustado, Y', ya arroja una primera interpretación: cuando permanece constante X_2, por cada hora de estudio, la puntuación en matemáticas aumenta en promedio, 1,899 puntos, y cuando permanece constante X_1, por cada punto más en razonamiento abstracto, aumenta 1,587 la puntuación e matemáticas

8.4.2.- Ajuste del modelo. Medidas de asociación

En regresión simple, el ajuste del modelo viene dado por el coeficiente de determinación que es el cuadrado del coeficiente de correlación de Pearson entre la VD y la VI, y ese coeficiente informaba de qué porción de la variabilidad de la VD es explicada por, o atribuida a, la variabilidad de la VI. En el caso de la regresión múltiple, las preguntas básicas que hay que responder son las siguientes:

- ¿Estiman bien la VD el **conjunto** de VI's?
- ¿Cuánta variabilidad explica cada variable individualmente una vez que las otras variables han aportado lo suyo?

Comencemos por responder a la primera pregunta, y para ello disponemos del denominado **coeficiente de correlación múltiple**, R, y su cuadrado, el **coeficiente de determinación**, R^2. Al igual que r es el coeficiente de correlación entre dos variables, R es el coeficiente que correlaciona la VD con una combinación óptima de dos o más variables independientes. Su fórmula de cálculo es:

$$R_{Y.12} = \sqrt{\frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}} \tag{8.28}$$

Y de forma alternativa, una expresión si cabe más simple es la basada en los coeficientes de regresión parcial estandarizados de la recta de regresión:

$$R_{Y.12} = \sqrt{\beta_1 r_{Y1} + \beta_2 r_{Y2}} \tag{8.29}$$

Aplicada la fórmula a los datos del ejemplo, el valor de $R_{Y.12}$ es:

$$R_{Y.12} = \sqrt{(0{,}469)(0{,}441) + 0{,}649(0{,}628)} = 0{,}7836$$

El coeficiente de determinación es el cuadrado del coeficiente de correlación múltiple, y su interpretación y cálculo es idéntica a la de la regresión simple: razón entre la varianza de los pronósticos y la varianza de la VD.

$$R^2_{Y.12} = \frac{S^2_{Y.12}}{S^2_Y} = (R_{Y.12})^2 \qquad (8.30)$$

En la Tabla 8.9 se muestran los valores de Y, los pronósticos y los residuos para los datos del ejemplo, cuya función de estimación de Y, ya calculada, es:

$$Y' = 1{,}899X_1 + 1{,}587X_2 + 5{,}217$$

Tabla 8.7

Puntuación en Matemáticas actual, estimada y residual para cada sujeto

	Punt. Matemáticas (Y)	Estimaciones (Y'_{12})	Residuos ($Y - Y'_{12}$)
	54	50,562	3,438
	52	50,874	1,126
	34	38,829	-4,829
	63	60,396	2,604
	46	52,461	-6,461
	44	47,7	-3,7
	50	54,984	-4,984
	52	44,526	7,474
	57	53,112	3,888
	53	59,433	-6,433
	56	51,186	4,814
	67	60,057	6,943
	57	60,396	-3,396
	54	52,461	1,539
	51	53,085	-2,085
Varianza	56,222	34,531	21,697

A partir de los datos de la Tabla 8.9 se obtiene el coeficiente de determinación R^2.

$$R^2_{Y.12} = \frac{S^2_{Y\cdot 12}}{S^2_Y} = \frac{34{,}531}{56{,}222} = 0{,}614$$

Es decir, la combinación de las dos variables (tiempo de estudio y razonamiento abstracto) se atribuyen el 61,4% de la variabilidad de las puntuaciones obtenidas en matemáticas, y por tanto el 38,6% restante se debe a otros factores no relacionados linealmente con dichas puntuaciones. Vemos que se cumple lo que denominamos **Teorema de Pitágoras de la Regresión Lineal**: la varianza de las puntuaciones observadas es igual a la varianza de las puntuaciones estimadas más la varianza de los residuos. En este caso, tomando los valores de las varianzas calculadas: 60,238 = 36,991 + 23,247.

El coeficiente R^2 obtenido en la muestra no es un estimador insesgado de ρ^2 en la población. Para entender esto de forma intuitiva, podemos imaginar el caso en que una o más VI's no contribuyen a la explicación de la varianza de la VD en la población. Sin embargo, en la muestra, debido a las fluctuaciones del proceso de muestreo, raramente se observa una situación en la que no haya contribución de una VI a la varianza de la VD, aunque sea muy pequeña. Cuanto menor sea la muestra mayor será la contribución a la VD, lo que provoca un aumento "artificial" de la R^2, valor que no se correspondería con el ρ^2 en la población. Por esa razón, es preferible disponer de una estimación más ajustada y realista de ρ^2. Este ajuste, se conoce como $\boldsymbol{R^2}$ **Ajustado** que simbolizaremos mediante la R mayúscula a la que se le superpone el signo virgulilla:

$$\tilde{R}^2_{Y.12} = 1 - (1 - R^2_{Y.12})\frac{n-1}{n-p-1} \qquad (8.31)$$

siendo *n*, el número de observaciones y *p*, el número de variables independientes o predictoras. Para el caso de ejemplo, el valor de R^2 Ajustado es:

$$\tilde{R}^2_{Y.12} = 1 - (1 - 0{,}614)\frac{15-1}{15-2-1} = 0{,}5498$$

Otro valor que informa del ajuste es el Error Típico (ya explicado para el caso bivariado) y que está relacionado con R^2 en el sentido de que cuando éste aumenta el Error Típico disminuye. De acuerdo a la ecuación 8.14, y siendo las sumas de cuadrados las que se muestran en la Tabla 8.10, su valor para este ejemplo es:

$$\sigma_\varepsilon = \sqrt{\frac{\sum(Y - Y')^2}{n-p-1}} = \sqrt{\frac{325{,}451}{15-2-1}} = 5{,}2078$$

Tabla 8.8

Sumas de cuadrados total, residual y debidas a la regresión del ejemplo numérico:

$$\sum (Y - \bar{Y})^2 = \mathrm{SC}_{\mathrm{Total}} = 843{,}333$$

$$\sum (Y - Y')^2 = \mathrm{SC}_{\mathrm{Residuos}} = 325{,}451$$

$$\sum (Y' - \bar{Y})^2 = \mathrm{SC}_{\mathrm{Regresión}} = 517{,}968$$

8.4.3.- Correlación Semiparcial y Parcial

La segunda de las preguntas que hacíamos al comienzo del epígrafe anterior, es cómo determinar la contribución de cada variable independiente a la explicación de la dependiente. La respuesta a esta pregunta la proporciona la llamada **correlación semiparcial**, *sr*, y su cuadrado, sr^2. Antes de explicar qué son esas nuevas correlaciones que acaban de entrar en escena, piense el lector que cuando en un modelo intervienen más de dos variables, las correlaciones que se calculan entre las variables dos a dos, no son correlaciones "puras", en el sentido de que no miden relaciones entre esas dos variables al margen del influjo que las otras variables del modelo puedan tener sobre cada una de ellas. Estas correlaciones que se calculan entre dos variables (correlaciones bivariadas) se denominan **correlaciones de orden cero**, y a través del valor obtenido no se puede saber qué parte de la varianza de la VD es capaz de explicar independientemente cada una de las VI's, puesto que entre éstas también puede haber relación. Por lo tanto, para saber qué parte de la VD explica cada VI al margen de las otras VI's, es necesario eliminar el influjo que sobre cada VI tienen el resto de las VI's, para así poder determinar el influjo único que esa VI tiene sobre la VD. Esta relación entre cada VI y la VD habiendo eliminado el influjo del resto de las VI's sobre cada VI es lo que se llama **Coeficiente de Correlación Semiparcial**.

¿Cómo se calcula este coeficiente? Ya sabemos, por todo lo explicado hasta el momento, que en un modelo de regresión hay una proporción de varianza explicada y una proporción de varianza no explicada que es la varianza de los residuos. La varianza explicada lo es en función de una cierta combinación de las variables independientes; por consiguiente, si en un modelo, por ejemplo, con dos predictoras X_1 y X_2, se ajusta una regresión de la 1 sobre la 2, se extraen los residuos y, por último, los correlaciono con la VD, habré calculado el coeficiente de correlación semiparcial entre X_1 y la VD habiendo eliminado el influjo de X_2 sobre la VD. Por otra parte, si se ajusta una regresión simple entre X_2 y X_1 (obsérvese el cambio de subíndices en relación a la frase anterior), se extraen los residuos y éstos se correlacionan con la VD, habré calculado la correlación entre el predictor X_2 y la VD, habiendo eliminado el influjo de X_1 sobre la VD.

Para llevar a cabo este cálculo de los coeficientes de correlación semiparcial no es necesario proceder como hemos explicado en el párrafo anterior; hay fórmulas muy sencillas para ello, a partir de las correlaciones de orden cero.

$$sr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r_{12}^2}}$$
$$sr_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1 - r_{12}^2}} \qquad (8.32)$$

y elevando al cuadrado estos valores se tiene la contribución que cada VI tiene sobre la VD habiendo eliminado el influjo de las otras VI's. En la Figura 8.8 se observa gráficamente, mediante un Diagrama de Venn, estás contribuciones expresadas en forma de área compartida

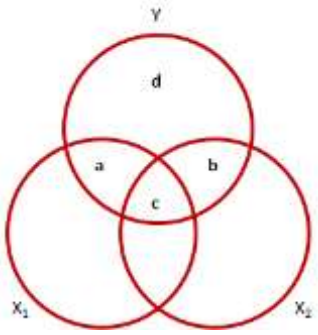

Figura 8.8 *Diagrama de Venn para un modelo de regresión con dos variables independientes*

Tomando como referencia el diagrama de la Figura 8.8, las equivalencias entre las zonas designadas con letras y los cuadrados de los coeficientes de correlación semiparcial, son las siguientes:

$$a = sr_1^2 = R_{Y.12}^2 - r_{Y2}^2$$
$$b = sr_2^2 = R_{Y.12}^2 - r_{Y1}^2 \qquad (8.33)$$

siendo:

$$R_{Y.12}^2 = a + b + c$$
$$r_{Y1}^2 = a + c$$
$$r_{Y2}^2 = b + c$$

Para el ejemplo numérico que sirve de base a la explicación, los cálculos de los coeficientes de correlación semiparcial son los siguientes:

$$sr_1 = \frac{0{,}4406 - (0{,}6285)(-0{,}0431)}{\sqrt{1-(-0{,}0431)^2}} = 0{,}4681$$

$$sr_2 = \frac{0{,}6285 - (0{,}4406)(-0{,}0431)}{\sqrt{1-(-0{,}0431)^2}} = 0{,}6481$$

Estos valores elevados al cuadrado dan la proporción de varianza compartida por cada predictora habiendo eliminado el influjo de la otra predictora sobre la misma.

$$a = sr_1^2 = 0{,}4681^2 = 0{,}2191$$

$$b = sr_2^2 = 0{,}6481^2 = 0{,}4200$$

El valor $0{,}4681^2$ (0,2191) es ***a*** en el diagrama de la Figura 8.8, y $0{,}6481^2$ (0,4200) es ***b***. Estos dos valores representan la contribución exclusiva que cada variable hace a la explicación de la dependiente. La porción ***c***, es la proporción de varianza de la VD estimada conjuntamente (es decir, de forma redundante) por las dos variables. Sin embargo esta proporción es de muy difícil interpretación.

El otro coeficiente que se calcula en los modelos de regresión, y que además sirve para determinar cuál es la primera variable que se incorpora al modelo cuando se realiza variable a variable[3], es el denominado **coeficiente de correlación parcial**, *pr*. La diferencia con el semiparcial es que en el parcial se elimina el influjo de los predictores tanto de la VI objeto de correlación como de la VD. Es decir, es una correlación entre residuos.

En el modelo de dos variables, si se ajusta una recta entre Y y X_2, y nos quedamos con los residuos, y si se ajusta una recta entre X_1 y X_2, y nos quedamos también con los residuos, podemos correlacionar ambos residuos. De esta forma obtendremos la correlación parcial entre Y y X_1. A partir de aquí se ve claro que esta es la correlación "pura" entre dos variables, puesto que de ambas se ha extraído el influjo de terceras variables. Al igual que en la correlación semiparcial, no es necesario el cálculo de los residuos, pues se pueden obtener a partir de los correlaciones de orden cero entre pares de variables.

$$pr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1-r_{Y2}^2}\sqrt{1-r_{12}^2}}$$

$$pr_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1-r_{Y1}^2}\sqrt{1-r_{12}^2}} \qquad (8.34)$$

El cuadrado de estos coeficientes (p.e. pr_1) se interpreta como la proporción de la varianza de la VD (Y) no asociada con X_2 que sí está asociada a X_1.

[3] Hay varios métodos para la introducción de variables en el análisis de regresión. Uno de estos métodos es el denominado *Stepwise* (Pasos Sucesivos) y en él se introduce en primer lugar la variable con mayor correlación con el criterio, y a partir de ahí, sucesivamente la variable que mayor correlación parcial tenga con el criterio. El proceso de introducción de variable se detiene cuando la siguiente variable independiente que va a entrar no aporta un plus significativo a la explicación de la VD.

Otra manera de calcular esta proporción de varianza es por medio de las porciones representadas en el diagrama de Venn de la Figura 8.8.

$$pr_1^2 = \frac{a}{a+d} = \frac{R_{Y.12}^2 - r_{Y2}^2}{1 - r_{Y2}^2}$$
$$pr_2^2 = \frac{b}{b+d} = \frac{R_{Y.12}^2 - r_{Y1}^2}{1 - r_{Y1}^2} \quad (8.35)$$

Aplicando las fórmulas a los datos del ejemplo, los coeficientes son:

$$pr_1 = \frac{0{,}441 - (0{,}628)(-0{,}043)}{\sqrt{1 - 0{,}628^2}\sqrt{1 - (-0{,}043)^2}} = 0{,}6018 \;\rightarrow pr_1^2 = 0{,}6018^2 = 0{,}3622$$

$$pr_2 = \frac{0{,}628 - (0{,}441)(-0{,}043)}{\sqrt{1 - 0{,}441^2}\sqrt{1 - (-0{,}043)^2}} = 0{,}7219 \;\rightarrow pr_2^2 = 0{,}7219^2 = 0{,}5211$$

Si se hubiera realizado una regresión paso a paso, es decir, introduciendo las variables por su relación con la VD, la primera que habría entrado en el modelo hubiera sido la variable X_2 (en el ejemplo, Razonamiento abstracto) que es la que presenta mayor correlación con la VD.

En resumen, por los resultados del coeficiente de correlación parcial y semiparcial al cuadrado, en el modelo obtenido está clara la contribución de ambas variables a la explicación de la puntuación en matemáticas. El cuadrado de los coeficientes *pr* señala la proporción de varianza de una VI asociada con la parte de la VD que no está asociada con la otra VI. En nuestro caso es mayor la de razonamiento abstracto que la de tiempo de estudio (52,11% y 36,22%, respectivamente). Además, el modelo es bueno (luego veremos su significación estadística, por medio de los contrastes) porque ambas variables independientes tienen una buena relación con la dependiente, y sin embargo, entre ellas no hay apenas relación (es, pues, un modelo casi ideal[4]). ¿Cómo se manifiesta numéricamente la ausencia de relación entre las variables independientes?, pues sencillamente en que el coeficiente de determinación, R^2 (0,6141), tiene un valor aproximado (siempre menor) que la suma de los cuadrados de los coeficientes de correlación semiparcial (0,2191+0,4200 = 0,6391 < 0.6141). La diferencia entre ambos valores es la parte redundante del diagrama de Venn (zona *c*) que el modelo de regresión elimina cuando se ajusta con el conjunto completo de variables independientes.

[4] Los datos del ejemplo son ficticios y han sido simulados para lograr este efecto de correlación media-alta de las variables predictoras con la VD y ausencia de correlación entre las predictoras. En análisis de regresión, cuando las VI's correlacionan se dice que hay "colinealidad", y cuanto mayor es ésta peor es el modelo de regresión.

8.5.- Resumen

El análisis de los diseños *ex post facto* trata de determinar cómo un conjunto de variables, que llamamos independientes, predictoras o explicativas, pueden explicar el comportamiento de la variable objeto de estudio, que llamamos dependiente o criterio. Ello se ha realizado en tres pasos:

- Ajuste del modelo de regresión para estimar la VD. Sólo se han tratado ajustes de modelo lineales, es decir, modelos en que la VD es una función lineal de la o las VI's. Cuando sólo hay una VI, el modelo se conoce como de **Regresión Lineal Simple** y cuando hay varias VI's, como de Regresión Lineal Múltiple.
- Cálculo de la bondad del modelo ajustado. El estadístico que cuantifica el ajuste se denominado **coeficiente de determinación** y su valor oscila entre 0 y 1, e informa de la proporción en que la o las VI's explican la VD. En el caso de la regresión simple, este valor es el cuadrado del coeficiente de correlación de Pearson, y en el caso de la regresión múltiple este valor es el cuadrado del coeficiente de correlación múltiple. La parte no explicada por el modelo de regresión es aquella que no está relacionada linealmente con la VD.
- Contraste de significación de los estadísticos del modelo en el caso de la regresión lineal simple.

Los diferentes coeficientes que han aparecido en el capítulo son:

- R, que expresa la correlación entre la VD (Y) y la mejor función lineal de las VI's (X_i's)
- R^2, que se interpreta como la proporción de varianza de VD asociada a la combinación lineal de las VI's. También se interpreta como la reducción proporcional del error inicial de la VD cuando se ajusta un modelo de estimación con las VI's.
- sr_i, coeficiente de correlación semiparcial, expresa la correlación entre Y y X_i, cuando de ésta se ha extraído la que mantiene con el resto de X_i's.
- sr_i^2, proporción de varianza de Y asociada únicamente la varianza de X_i, y expresa el incremento en R^2 cuando la variable X_i entra en el modelo
- pr_i, expresa la correlación "pura" entre Y y X_i. Es decir, expresa la correlación entre la parte de Y no asociada linealmente con el resto de predictoras y la porción de X_i no asociada linealmente con el resto de predictoras.
- pr_i^2, expresa la proporción de varianza de Y no asociada al resto de X que sí está asociada con X_i.

8.6.- Ejercicio de Autoevaluación

Todas las preguntas están relacionadas con datos de una investigación (ficticia, con datos simulados) en la que se trata de determinar la influencia que sobre el resultado en las pruebas para acceder a un puesto de trabajo especializado tienen una serie de variables, como son los días que asisten a tutoría en una escuela de formación para ese tipo de profesionales (variable X_1), y la expectativa de empleo que manifiestan los sujetos (variable X_2), variables todas ellas cuantitativas o métricas. Como variable dependiente se toma, como se ha señalado, el resultado en una prueba en términos de puntuación obtenida (variable *Y*). Los datos de 25 personas son los siguientes:

X_1	X_2	Y
31	9	108
41	6	86
20	9	80
41	7	79
40	9	96
28	9	79
41	9	98
37	8	86
41	6	89
39	11	92
56	9	111
43	11	102
42	10	89
36	7	90
36	13	112
32	7	83
49	8	104
45	11	98
20	10	88
33	11	106
39	13	110
19	10	92
27	12	92
17	11	81
29	13	103

Para facilitar los cálculo, en las siguientes dos tablas presentamos los estadísticos descriptivos de cada variable, y la matriz de correlaciones

Estadísticos descriptivos

	X_1	X_2	Y
Suma	882	239	2354
Media	35,2800	9,5600	94,1600
Desv. Típica	9,5143	2,0412	10,3293
Varianza	90,5216	4,1664	106,6944

Matriz de correlaciones de orden cero

	X_1	X_2	Y
X_1		-0,231	0,436
X_2			0,504
Y			

Preguntas

1. ¿Cuál es la ecuación de regresión para la predecir el comportamiento de la variable Y a partir de la variable X_1?
    a. $Y' = 77{,}465 + 0{,}473X_1$
    b. $Y' = 35{,}465 + 0{,}573X_1$
    c. $Y' = 77{,}465 + 0{,}743X_1$
2. ¿Cuál es la ecuación de regresión para la predecir el comportamiento de la variable Y a partir de la variable X_2?
    a. $Y' = 44{,}236 + 1{,}873X_2$
    b. $Y' = 69{,}768 + 2{,}551X_2$
    c. $Y' = 77{,}465 + 0{,}743X_1$
3. El coeficiente de correlación múltiple del modelo $Y' = B_0 + B_1X_1 + B_2X_2$ para los datos propuestos es:
    a. 0,874
    b. 0,759
    c. 0,576
4. El coeficiente R^2 ajustado para los datos es:
    a. 0,594
    b. 0,512
    c. 0,538
5. Siguiendo el método de Pasos Sucesivos (Stepwise) para lograr el mejor ajuste, ¿qué cambio se produce en R^2 cuando se incorpora la segunda variable?
    a. 0,322
    b. 0,254
    c. 0,222

6. La ecuación de regresión múltiple estandarizada para los datos es:
   a. $z'_y = 0{,}423z_1 + 1{,}436z_2$
   b. $z'_y = 1{,}014z_1 + 0{,}872z_2$
   c. $z'_y = 0{,}583z_1 + 0{,}639z_2$
7. La varianza de los errores una vez ajustado el modelo de regresión múltiple es:
   a. 47,109
   b. 64,031
   c. 111,140
8. El error típico de estimación del modelo ajustado es:
   a. 7,891
   b. 7,169
   c. 8,235
9. La correlación entre la variable dependiente Y y la predictora X_1, una vez que se ha eliminado el influjo de X_2 sobre ambas variables, es:
   a. 0,659
   b. 0,567
   c. 0,621
10. ¿Cuál es la proporción de la varianza de Y asociada a X_2, y no asociada a X_1
   a. 0,234
   b. 0,342
   c. 0,477

Solución ejercicios de autoevaluación

Debajo de las respuestas están las operaciones necesarias, a partir de los estadísticos y la matriz de correlaciones.

Pregunta 1 A

Pregunta 2 B

$$B_1 = r_{Y1}\frac{S_Y}{S_{X_1}} = 0{,}436\frac{10{,}5423}{9{,}7105} = 0{,}473$$

$$B_0 = \bar{Y} - B_1\bar{X}_1 = 94{,}16 - (0{,}473)(35{,}28) = 77{,}465$$

$$B_1 = r_{Y2}\frac{S_Y}{S_{X_2}} = 0{,}504\frac{10{,}5423}{2{,}0833} = 2{,}5514$$

$$B_0 = \bar{Y} - B_1\bar{X}_2 = 94{,}16 - (2{,}5514)(9{,}56) = 69{,}768$$

Pregunta 3. B

$$R_{Y.12} = \sqrt{\frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}} = \sqrt{\frac{0{,}436^2 + 0{,}504^2 - 2(0{,}436)(0{,}504)(-0{,}231)}{1 - (-0{,}231)^2}} = 0{,}759$$

Pregunta 4. C

$$\tilde{R}^2_{Y.12} = 1 - \left(1 - R^2_{y.12}\right)\frac{n-1}{n-p-1} = 1 - (1 - 0{,}759^2)\frac{25-1}{25-2-1} = 0{,}538$$

Pregunta 5. A

$$R^2_{Y.12} - r^2_{Y2} = 0{,}759^2 - 0{,}504^2 = 0{,}322$$

El método *Stepwise*, la primera variable en entrar en el modelo sería la X_2 pues es la que más correlaciona con Y

Pregunta 6. C

$$\beta_1 = \frac{r_{Y1} - r_{y2}r_{12}}{1 - r^2_{12}} = \frac{0{,}436 - (0{,}504)(-0{,}231)}{1 - (-0{,}231)^2} = 0{,}583$$

$$\beta_2 = \frac{r_{Y2} - r_{y1}r_{12}}{1 - r^2_{12}} = \frac{0{,}504 - (0{,}436)(-0{,}231)}{1 - (-0{,}231)^2} = 0{,}639$$

Pregunta 7. A

$$S^2_{Error} = (1 - R^2_{Y.12})S^2_Y = (1 - 0{,}759^2)(111{,}14) = 47{,}109$$

Pregunta 8. B

$$\sigma_\varepsilon = \sqrt{\frac{\sum(Y - Y`)^2}{n-p-1}} = \sqrt{\frac{1130{,}6}{25-2-1}} = 7{,}169$$

El numerador del cociente dentro de la raíz es la suma de cuadrados de los errores, y se obtienen mediante

$$\sum(Y - Y`)^2 = S^2_{Error}(n-1) = (47{,}109)(25-1) = 1130{,}6$$

Pregunta 9. A

Se trata del coeficiente de correlación parcial entre las variable Y y X_1.

$$pr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r^2_{Y2}}\sqrt{1 - r^2_{12}}} = \frac{0{,}436 - (0{,}504)(-0{,}231)}{\sqrt{1 - (0{,}504)^2}\sqrt{1 - (-0{,}231)^2}} = 0{,}657$$

Pregunta 10. C

$$pr^2_2 = \left(\frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1 - r^2_{Y1}}\sqrt{1 - r^2_{12}}}\right)^2 = \left(\frac{0{,}504 - (0{,}436)(-0{,}231)}{\sqrt{1 - (0{,}436)^2}\sqrt{1 - (-0{,}231)^2}}\right)^2 = 0{,}477$$