2

Regresión lineal múltiple

1. Introducción.

En el tema anterior estudiamos la correlación entre dos variables y las predicciones que pueden hacerse de una de ellas a partir del conocimiento de los valores de la otra, es decir, se pronosticaban valores determinados de una variable criterio (Y) en función de según qué valores se obtenían de una variable predictora (X).

Sin embargo, la predicción de una variable (Y) a partir del conocimiento de otra única puede resultar un análisis extremadamente simple de la realidad en la cual existen múltiples factores que afectan a cualquier fenómeno que pretenda explicarse. En Psicología diremos que la conducta de los sujetos, en sus diversas manifestaciones, constituye función de múltiples variables que la condicionan y/o determinan. Por ejemplo, el comportamiento de un sujeto en una situación conflictiva puede depender de su temperamento, del nivel de conflictividad de la situación, de su experiencia en situaciones de este tipo por las que haya pasado previamente, etc... Es importante, pues, estudiar con un modelo de análisis más complejo que el de la regresión simple, de qué manera se producen estas relaciones entre la conducta y una serie más o menos numerosa de factores que la condicionan. El modelo de la regresión múltiple va a permitirnos acometer esta tarea.

1.1.Ecuación de regresión múltiple y supuestos.

Centrémonos, por ejemplo, en la variable éxito académico. Intentemos identificar las variables de las que depende la calificación final de un sujeto que ha cursado la enseñanza secundaria y se presenta inminentemente a la selectividad. Explicar de forma exclusiva dicha nota a partir del nivel de inteligencia de los sujetos puede resultar demasiado simplista e incompleto. Posibles condicionantes adicionales que influyen sobre la misma pueden ser: el nivel social o contexto cultural en el que se desenvuelve el sujeto, la motivación que siente por el estudio, el tiempo que emplea en dicho estudio o la calidad de los profesores que haya tenido en su trayectoria estudiantil...

Seleccionemos, de todas estas, dos variables para confeccionar un modelo explicativo del éxito académico (la nota en una prueba final): La inteligencia y la motivación por el estudio. Formalizamos este modelo, de manera similar a como lo hacíamos en regresión simple, con una ecuación del tipo:

$$Y = a + b_1 X_1 + b_2 X_2 + e$$

donde X_1 y X_2 son las variables predictoras contempladas, inteligencia y motivación, respectivamente. Sus coeficientes correspondientes representan el cambio previsto en Y por cada unidad de cambio en cada X, manteniendo la otra variable X constante. Así

pues, en el caso de X_1 , su coeficiente b_1 denota el cambio esperado en el éxito académico –calificación- por cada punto más en el variable inteligencia bajo un determinado valor de la motivación; es decir, sin tener en cuenta a esta segunda variable. El término e representa el error de predicción del modelo.

En este sentido, supongamos que en el ejemplo planteado el valor de los diferentes coeficientes fuera:

$$Y = 2 + 0.5X_1 + 2X_2 + e$$

Interpretaríamos a cada uno de ellos así:
a: En el caso de poseer una inteligencia y un nivel de motivación nulos para los objetivos de aprendizaje en cuestión, la calificación final esperada del sujeto sería 2.
X_1: Independientemente del nivel de motivación del sujeto por el aprendizaje, por cada unidad sumada a la variable inteligencia se incrementa 0.5 puntos la nota en la prueba final.
X_2: La nota esperada se verá incrementada en 2 puntos por cada unidad de cambio en la variable motivación, sea cual sea el nivel de inteligencia del sujeto.

Cada variable X mantiene con la variable criterio Y una relación de *linealidad*, es decir, supone un incremento constante y regular en Y por cada cambio en X. En el modelo global el cambio total de Y se debe a la suma de los incrementos por separado de cada variable predictora. En definitiva, el modelo se considera aditivo en este sentido descrito por ser una suma de efectos.

Para representar gráficamente el modelo de regresión con dos variables predictoras se necesita un espacio de tres dimensiones (como en la regresión simple se necesitaban dos dimensiones), es decir, una dimensión para cada una de las variables del modelo: dos para cada una de las Xs, respectivamente, y otra más para la variable Y. En esta circunstancia se obtiene un plano de regresión (no una recta como en regresión simple). El lugar donde dicho plano corta el eje de la variable Y es su ordenada en el origen, es decir, el valor de Y cuando tanto X_1 como X_2 valen 0. El desplazamiento en el eje de Y en función de X_1 refleja los cambios de X_1 sobre Y manteniendo constante X_2 . Y de forma similar, el desplazamiento en el eje de Y respecto a X_2 refleja la relación entre X_2 e Y manteniendo constante X_1.

Además, al igual que en el modelo de regresión simple, otros requisitos que deben cumplir los datos son:

- *Homocedasticidad*: La distribución de los errores respecto al plano de regresión es constante, es decir, homogénea alrededor del mismo

- *Normalidad*: Dicha distribución de errores sigue una ley normal

- *Independencia de errores:* Los errores son independientes entre sí, no están relacionados tampoco con las variables predictoras ni con la criterio. En suma, las puntuaciones de Xs e Y no se influyen unas con otras.

Además de estas condiciones apuntadas, para establecer adecuadas estimaciones de los coeficientes de la ecuación, el modelo de regresión múltiple requiere que las variables predictoras no presenten entre sí correlaciones altas. Si se dan correlaciones altas entre

ellas, estos coeficientes (b_1, b_2,...) pueden sufrir grandes cambios debido a que cada uno de ellos refleja el efecto específico de cada predictor con el criterio eliminando en dicho cálculo la correlación que dicho predictor mantiene con el resto de predictores así como la que mantienen estos predictores con el criterio. Así, existiendo una correlación entre dos predictores, el cálculo del coeficiente parcial del primero será reflejo de aquélla correlación exclusiva entre éste y el criterio eliminando de dicho cálculo tanto la supuesta relación que la segunda variable mantenga con el criterio como la correlación entre el primer y segundo predictor. En este sentido se dice que los coeficientes de la ecuación múltiple son coeficientes de *correlación parcial* por lo que la existencia de grandes porcentajes de variabilidad compartida entre predictores –colinealidad- hace que las estimaciones calculadas sufran muchos cambios respecto a las estimadas independientemente entre cada regresor particular con la variable criterio. Así pues, las interpretaciones teóricas en estas situaciones de existencia de alta colinealidad resultan muy difíciles.

1.2. Coeficiente de regresión múltiple y R^2 múltiple.

Se define el coeficiente de regresión múltiple como la correlación existente entre la variable criterio –Y- y el conjunto de las variables predictoras contempladas en el modelo. A diferencia del coeficiente de correlación simple, el coeficiente de correlación múltiple es siempre positivo por lo que la naturaleza de la relación de cada predictor (positiva o negativa) con la variable criterio no se refleja en el resultado. Si queremos conocer el signo que determina la relación de cada variable predictora con el criterio debemos identificar el signo que acompaña a su coeficiente de regresión particular en la ecuación de regresión múltiple o calcular su coeficiente de correlación simple (bivariado) con la variable criterio.

El cuadrado del coeficiente de correlación múltiple representa la proporción de la variabilidad de Y explicada por el conjunto de las Xs, es decir por el componente explicativo, conocido o determinista del modelo. Como complemento, 1- R^2 constituye como sabemos la proporción de variación no explicada o residual atribuida al efecto de factores aleatorios y desconocidos, ajenos a las variables predictoras analizadas.

Tal y como sabemos -y también para el modelo de regresión múltiple-:

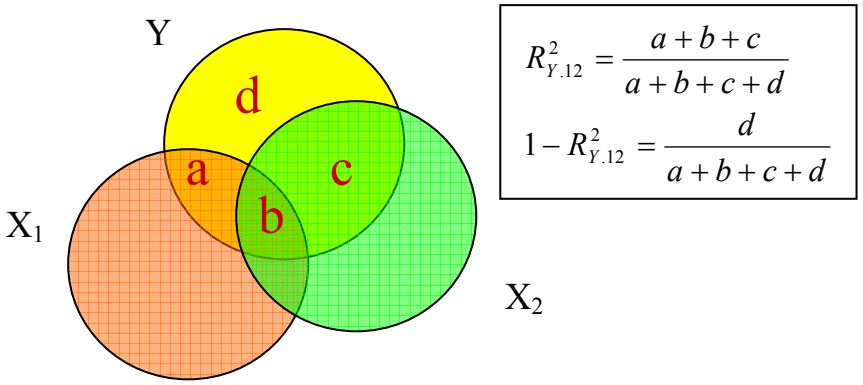
$$R^2 = \frac{SC_{regresión}}{SC_{total}} = \frac{\sum_{1}^{N}(\hat{Y} - \overline{Y})^2}{\sum_{1}^{N}(Y - \overline{Y})^2}$$

es decir, la proporción que de la variabilidad de los datos de Y respecto a su media (variabilidad total –en el denominador-) se atribuye a la regresión (variabilidad explicada –en el numerador-).

Entonces la proporción de variación no explicada es:

$$1 - R^2 = \frac{SC_{residual}}{SC_{total}} = \frac{\sum_{1}^{N}(Y - \hat{Y})^2}{\sum_{1}^{N}(Y - \overline{Y})^2}$$

Resulta muy útil, para interpretar estos valores de variabilidad explicada y residual del modelo de regresión múltiple, recurrir a procedimientos gráficos utilizando diagramas de Venn, donde cada variable está representada por un círculo. Supongamos que tratamos de explicar Y por cuenta de dos variables predictoras (X_1 y X_2). Las diferentes porciones de variabilidad de Y compartidas ($R^2_{Y.12}$) y no compartidas ($1-R^2_{Y.12}$) con estas variables X son:



En este diagrama puede observarse que la parte compartida de Y con X_1 y X_2 viene determinada por el área total a + b + c y la parte no explicada por el área d. Las proporciones de variación explicada y no explicada de este modelo planteado son las que se determinan arriba en el cuadro ($R^2_{Y.12}$ y $1-R^2_{Y.12}$). Así:

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

1.3. Validación del modelo.

Como se sabe, validar un modelo de regresión consiste en analizar si la variabilidad de la variable criterio (Y) atribuida a la regresión –en este caso al efecto del conjunto de variables predictoras- es lo suficientemente grande con respecto a la variabilidad no explicada o residual. El índice F constituye una prueba estadística pertinente para evaluar dicha relación:

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

La probabilidad (p) asociada al resultado de dicha prueba indica el grado de certidumbre con el que podemos concluir que numerador -parte explicativa del modelo- y denominador -parte borrosa o residual- coinciden, es decir, que lo determinado o explicativo se confunde con –o es lo mismo a- lo borroso del modelo. Si dicha probabilidad es pequeña ($p<.05$) concluimos que la parte explicativa supera en cantidad suficiente a la no explicada, por lo que las variables determinadas como relevantes por el modelo se consideran significativas –en su conjunto-.

Es importante anotar que validar de esta forma un modelo en regresión múltiple significa que el modelo en su conjunto lo es, es decir, el conjunto de predictores

contemplados logran explicar una porción importante de la variabilidad de Y; sin embargo mediante dicha prueba no se sabe nada sobre el poder explicativo de cada predictor por separado. Puede ocurrir que el modelo en su conjunto tenga un poder de explicación alto, sin embargo alguna de las variables predictoras no lo tenga, es decir, no sea significativa su relación con Y. El investigador debe depurar el modelo planteado de cara a eliminar variables insignificantes o claramente redundantes con otras del modelo a fin definir el modelo más parsimonioso posible respecto a la realidad que pretende explicar.

Además hay que tener en cuenta que eliminar o añadir variables al modelo no sólo hace reducir o agrandar su poder explicativo sino que puede afectar drásticamente a los coeficientes que acompañan a los restantes predictores y por ello a sus poderes explicativos (recuérdese que los coeficientes de estos predictores son coeficientes de correlación parcial). Estos cambios se producen tanto más drásticamente cuanto mayor correlación exista entre las variables eliminadas o introducidas en el modelo y las que se mantienen en él.

1.4. Depuración del modelo.

Cuando se estima un modelo de regresión múltiple con un número determinado de variables predictoras (Xs), uno de los procedimientos aplicados para estimar la ecuación de regresión múltiple correspondiente es el de introducir por igual y simultáneamente todos los predictores deseados en dicho modelo. En el paquete estadístico SPSS dicho procedimiento es el denominado "método introducir". Supuestamente este procedimiento resulta especialmente útil en los casos en que el investigador no tenga de antemano una idea jerárquica de la importancia relativa de cada variable predictora, lo que quizás vaya acompañado de un escaso conocimiento teórico sobre el tema bajo estudio. En este caso se estima el modelo de regresión múltiple completo y si a posteriori se aprecia que alguna o algunas de las variables no mantienen con el criterio relación significativa, se van eliminando en pasos posteriores de cara a depurar dicho modelo.

Otro manera de estimar el modelo de regresión múltiple, más sofisticada, es el método "*stepwise*" (en el SPSS) o método de estimación por pasos. En su modalidad "fordward", consiste en ir estimado sucesivos modelos de regresión, según se incorporen, una a una y paso a paso, las diferentes variables predictoras que se suponen relevantes para explicar Y. El orden en que van incorporándose dichas variables predictoras en estos sucesivos modelos es determinado por el investigador según prioridades teóricas o de otra índole.

Mediante este método de estimación por pasos no sólo se evalúa la significación de cada modelo estimado (con una o más variables predictoras) sino que se informa también del aumento del poder explicativo de dicho modelo según van incorporándose, una a una, las restantes variables potencialmente explicativas. Dicho poder explicativo tiene que ver con el denominado *coeficiente de correlación semiparcial* de cada una de estas variables -paso a paso introducidas- con la variable criterio. Lo desarrollaremos posteriormente.

2. Aplicación práctica de un ejemplo de regresión múltiple.

Supongamos que obtenemos los siguientes datos en el estudio de la relación citada sobre la inteligencia y la motivación como predictores de la calificación final obtenida en el bachillerato en una muestra de 12 sujetos:

Sujetos	X1 (Inteligencia)	X2 (motivación)	Y (Nota)
1	85	10	4
2	100	20	5
3	95	35	8
4	80	30	7
5	180	45	10
6	90	25	6
7	110	10	6
8	120	15	7
9	80	10	4
10	95	15	4
11	160	15	6
12	150	45	9
Medias	112.08	22.92	6.33
S	33.60	13.05	1.97

2.1. Ecuación del modelo y bondad de ajuste.

Para estimar la ecuación de regresión múltiple para estos datos en el paquete estadístico SPSS se procede de forma similar a como se hacía en el modelo de regresión simple. La sucesión de comandos es: Analizar/Regresión/Lineal especificando que la variable dependiente es la nota (Y) y las variables independientes la inteligencia (X_1) y la motivación (X_2). Por defecto, estimamos el modelo mediante el método "introducir", mediante el que, como hemos explicado, todas las variables predictoras se introducen a la vez en el modelo estimado. Algunos resultados de la salida proporcionada son los siguientes:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,922(a)	,849	,816	,84514

a Variables predictoras: (Constante), motivación, Inteligencia

Tales resultados indican que el coeficiente de correlación múltiple es .922, es decir, la correlación entre el conjunto de variables predictoras (X_1 y X_2) y la criterio (Y), por lo que el 84.9% de la variabilidad de Y ($.922^2$) se explica por las variables predictoras contempladas en el modelo. El ajuste del modelo puede considerarse según este dato bastante alto.

El coeficiente de correlación múltiple al cuadrado corregido es una forma de minimizar la suma de todas las correlaciones que mantienen la serie de variables predictoras con el criterio; es decir, una forma de aminorar lo inflado que puede resultar dicho coeficiente. Se calcula así:

$$R^2_{ajus} = 1 - (1 - R^2)\frac{(N-1)}{(N-k-1)}$$

siendo, pues, una función del número de sujetos de la muestra y del número de predictores contemplados en la ecuación; por tanto será mayormente minimizado cuanto mayor es N y mayor el número de predictores contemplados. En nuestro caso:

$$R^2_{ajus} = 1 - (1 - .849)\frac{(12-1)}{(12-2-1)} = .816$$

Por último, el error típico de estimación se refiere a la desviación típica de las puntuaciones de error, es decir, a la raíz cuadrada de la varianza residual:

$$S_{res} = \sqrt{\frac{\sum_1^N (Y - \hat{Y})^2}{N-k-1}}$$

Por otro lado, el siguiente cuadro nos proporciona información sobre la ecuación de regresión y sus coeficientes:

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	1,737	,885		1,964	,081
	Inteligencia	,019	,009	,320	2,153	,060
	motivación	,109	,022	,721	4,847	,001

a Variable dependiente: nota

Según la información proporcionada, la ecuación que representa la regresión de la inteligencia (X_1) y la motivación (X_2) sobre la nota es:

$$\hat{Y} = 1.737 + 0.019X_1 + 0.109X_2$$

lo cual refleja que por un punto más en la variable inteligencia (X_1) se incrementa la nota sólo .019 puntos. Respecto a la motivación (X_2), la nota sube 10.9 décimas por cada unidad más puntuada en esta última variable[1]. Por último, la nota prevista en ausencia de motivación y de inteligencia alguna es de 1.737.

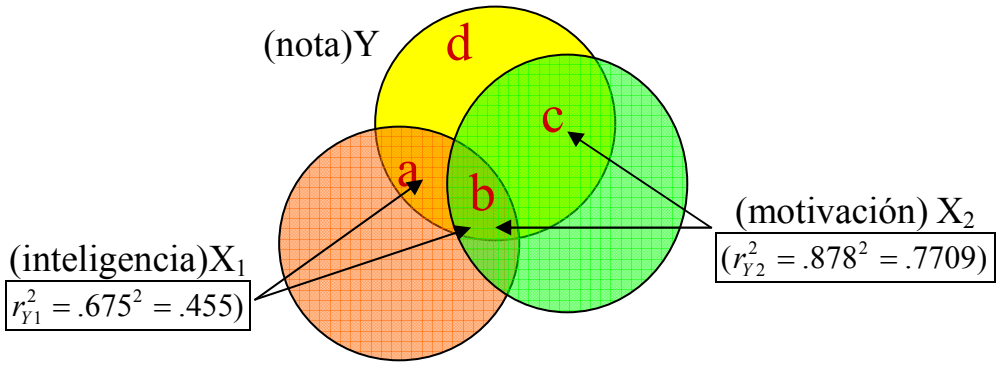
[1] Hay que tener en cuenta que las escalas en que se miden estas variables (inteligencia y motivación) son bastante diferentes en su amplitud (mayor en inteligencia que en motivación) por lo que es previsible hasta cierto punto que una unidad de cambio en la primera ejerza un efecto menor sobre la Y que una unidad de cambio en la segunda.

2.2. Coeficientes de correlación parcial y correlación parcial.

Como hemos explicado, los coeficientes obtenidos en la ecuación de regresión múltiple representan la contribución específica de cada variable X sobre Y con independencia de las restantes –o manteniéndolas constantes-. Es decir, el incremento de la nota en .019 puntos debido a la inteligencia se produce independientemente a la puntuación que exista en la variable motivación y a la inversa, el efecto sobre la nota en .109 puntos por cada punto más en motivación resulta para cualquier valor obtenido en la variable inteligencia.

Así pues, si en alguna medida las variables motivación e inteligencia están relacionadas -comparten variabilidad- el coeficiente asociado a la inteligencia proporciona información de su contribución específica sobre la nota eliminando aquella variación que comparte con la motivación. El valor de los coeficientes en la ecuación de regresión múltiple tiene que ver, como hemos dicho, con la denominada *correlación parcial* entre las variables. Así, en la ecuación de regresión múltiple la interpretación de cada coeficiente remite – o es proporcional- a su correlación parcial, por lo que podríamos decir, por ejemplo, que la correlación parcial entre la inteligencia y la nota es aquella correlación calculada entre inteligencia y nota en el caso de que todos los sujetos tuvieran idéntica motivación.

De esta forma los coeficientes de la ecuación de regresión múltiple representan, al hilo de lo expuesto, contribuciones de cada variable predictora sobre la criterio eliminando el efecto compartido con las restantes. Esto es, constituye un valor específico relacionado con la correlación de cada predictor (X) de forma exclusiva con el criterio (Y). Mediante un diagrama de Venn este concepto puede entenderse más fácilmente. Representemos -como anteriormente- a cada variable en la ecuación, mediante un círculo que denota su variabilidad total y ciertos solapamientos entre estos círculos que muestran la cantidad de información compartida -o correlación al cuadrado- entre ellas.



Observando el gráfico diríamos que:

1) (a+b)/(a+b+c+d) es el valor de la correlación bivariada al cuadrado entre inteligencia y nota ($r^2_{Y1} = .675^2 = .455$)
2) (b+c)/(b+c+a+d) es el valor de la correlación bivariada al cuadrado entre motivación y nota ($r^2_{Y2} = .878^2 = .7709$) .

3) (a)/(a+d) es el valor de la correlación parcial al cuadrado entre inteligencia y nota ($r^2_{Y1.2}$)
4) (c)/(c+d) es el valor de la correlación parcial al cuadrado entre motivación y nota ($r^2_{Y2.1}$)

Así, los coeficientes de correlación parcial estimados en la ecuación múltiple tienen que ver con las áreas a y c para inteligencia y motivación, respectivamente, y no con las áreas (a + b) y (b + c) que estarían relacionadas con los coeficientes estimados para cada una de estas variables en sendas ecuaciones de regresión simple –independientes- en su influencia sobre la Y.

De esta forma tenemos que, en su conjunto, la variable notas (Y) queda explicada por las dos variables introducidas en el modelo múltiple (inteligencia y motivación) en las porciones a + b + c, por lo que el coeficiente de correlación múltiple al cuadrado se calcula de esta forma como vimos antes:

$$R^2_{Y.12} = \frac{a+b+c}{a+b+c+d}.$$

Según quedaba calculado en la salida anteriormente expuesta del paquete estadístico SPSS, dicho coeficiente vale para el caso que tratamos:

$$R^2_{Y.12} = .849$$

¿Cómo se calcula la correlación parcial al cuadrado entre variables? Como hemos dicho, esta correlación parcial constituye una información relevante en el modelo de regresión múltiple porque refleja la cuantía de los coeficientes de regresión –que son parciales- en la ecuación de regresión múltiple.

Por ejemplo, reparemos primero en la correlación parcial al cuadrado entre X_1 con Y - $R^2_{Y1.2}$ -(por lo tanto eliminando de ella todo lo que se refiere a X_2) para nuestros datos. Como hemos dicho, en el diagrama de Venn dibujado quedaría reflejada en el área definida como "a" respecto a lo que queda de Y habiendo eliminado aquella porción que tiene que ver con X_2 (b+c). Es decir, la porción "a" respecto a "d". Formalmente:

$$R^2_{Y1.2} = \frac{a}{a+d} \quad o$$

$$R^2_{Y1.2} = \frac{R^2_{Y.12} - R^2_{Y2}}{1 - R^2_{Y2}}$$

Sustituyendo los valores correspondientes:

$$R^2_{Y1.2} = \frac{R^2_{Y.12} - R^2_{Y2}}{1 - R^2_{Y2}} = \frac{0.849 - 0.7709}{1 - 0.7709} = 0.3398 \quad ; \quad r_{Y1.2} = \sqrt{0.398} = 0.583$$

o bien, tomando como referencia las correlaciones simples entre cada par de variables, la expresión anterior puede quedar así:

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2} \cdot r_{12}}{\sqrt{1 - r_{Y2}^2} \cdot \sqrt{1 - r_{12}^2}} = \frac{0.675 - 0.878 \cdot 0.493}{\sqrt{1 - 0.878^2} \cdot \sqrt{1 - 0.493^2}} = 0.583$$

Desde el punto de vista analítico, pues, las correlaciones que presenta cada una de las variables predictoras con la variable criterio (correlaciones bivariadas) no se corresponden con las correlaciones de tipo parcial sobre las que se basa el cálculo de los coeficientes en la ecuación múltiple. Es decir, en las correlaciones bivaridas, de predictora y criterio entre sí, no queda excluida la posible correlación existente entre ellas y otra tercera variable.

Para confirmar esto que decimos basta analizar cuáles son los valores de las correlaciones bivariadas para nuestros datos:

Correlaciones

		Inteligencia	motivación	nota
Inteligencia	Correlación de Pearson	1	,493	,675(*)
	Sig. (bilateral)		,104	,016
	N	12	12	12
motivación	Correlación de Pearson	,493	1	,878(**)
	Sig. (bilateral)	,104		,000
	N	12	12	12
nota	Correlación de Pearson	,675(*)	,878(**)	1
	Sig. (bilateral)	,016	,000	
	N	12	12	12

Se observa que la variable inteligencia muestra con la nota una correlación de .675 (estadísticamente significativa al .05) lo cual supone que el 45% ($.675^2$ x 100) de la variabilidad de la nota se explica por el efecto de la inteligencia. ¿A qué es debido entonces que el coeficiente –parcial- asociado a dicha variable en la ecuación múltiple sea tan pequeño y no significativo (p=.06)? Pues precisamente a la correlación compartida entre inteligencia y motivación (según la tabla anterior igual a .493). Esta correlación (.493) resulta eliminada para estimar el coeficiente de correlación parcial de la inteligencia sobre la nota en la ecuación múltiple, lo que hace que el valor de la correlación restante –parcial- no resulte significativo. Dicho de otra forma: El coeficiente asociado a la relación entre inteligencia y nota se refiere al efecto exclusivo de la inteligencia sobre la nota una vez eliminada la correlación que ésta mantiene con la motivación. Así, podemos afirmar que cuanto mayor sea la correlación que mantienen entre sí dos predictores en una ecuación múltiple, más cambio se aprecia en los valores de los coeficientes de correlación parcial respecto a los coeficientes de correlación simple. Un análisis similar puede realizarse respecto a la variable motivación para comprobar que su correlación bivariada con las notas no coincide con su correlación parcial con la misma.

Si hubiésemos probado en una ecuación de regresión simple el efecto de la inteligencia sobre la nota los resultados hubieran sido:

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	T	Sig.
		B	Error típ.	Beta		
1	(Constante)	1,897	1,594		1,190	,261
	Inteligencia	,040	,014	,675	2,896	,016

a Variable dependiente: nota

un efecto estadísticamente significativo, que no lo es en la ecuación múltiple. Obsérvese la coincidencia del coeficiente estandarizado (Beta) con el valor de la correlación bivariada calculada en el cuadro anterior.

La interpretación de los coeficientes de la ecuación de regresión múltiple, pues, hay que hacerla con mucha cautela. Un coeficiente pequeño no es reflejo de una baja correlación entre la variable que lo acompaña y el criterio sino que puede significar que la información compartida entre dicha variable y otra (u otras) predictoras del modelo es muy alta. La recomendación es, teniendo en cuenta que el modelo se elabora con estos condicionantes, que las variables predictoras que explican Y sean lo más independientes entre sí, es decir, no mantengan relación o colinealidad . Sin embargo, la mayoría de las veces la ausencia total de colinealidad entre ellas es imposible de satisfacer y por tanto hay que jugar con dicha información compartida entre preeditores a la hora de aportar la interpretación última de la ecuación de regresión. Algunos autores (véase, por ejemplo, Gardner 2003) recomiendan, teniendo en cuenta este problema, que no es conveniente interpretar y evaluar la importancia de los predictores con base en los resultados únicos de una ecuación de regresión múltiple.

2.3. Coeficientes estandarizados.

¿Qué ventaja poseen los coeficientes de regresión estandarizados respecto a los recién interpretados no estandarizados?. En principio, la problemática expuesta sobre la interpretación de los coeficientes de regresión parcial en la ecuación múltiple es también aplicable para aquellos coeficientes estandarizados. Sin embargo, es importante considerar que estos últimos presentan una ventaja esencial sobre los coeficientes no estandarizados. Y es la posibilidad de poder evaluar y comparar el poder explicativo de cada predictor en la ecuación al ser directamente comparables, cosa imposible a través de los coeficientes directos que dependen de la escala en que se miden las diferentes variables. Los coeficientes estandarizados remiten a una escala única (en desviaciones típicas respecto al 0) en que se miden las diferentes variables y por tanto pueden constituir la base para conocer exactamente en cuántos puntos se modifica la variable Y por cuenta de cada regresor comparativamente. Recordemos que para nuestros datos la ecuación en puntuaciones directas era:

$$\hat{Y} = 1.737 + 0.019X_1 + 0.109X_2$$

En dicha ecuación no puede interpretarse que el efecto de la motivación (X_2) sobre la nota es aproximadamente 0.1 unidades mayor que el de la inteligencia (X_1) (0.109 – 0.019 = .09), ya que cada coeficiente está calculado a partir de la escala específica en que se mide su correspondiente variable: de 0 a 200 la inteligencia y de 0 a 50 la

motivación. En función de estas escalas el coeficiente asociado a la inteligencia, suponiéndole un efecto igual sobre la nota que la motivación, sería menor que el asociado a esta última variable, con tanta más diferencia cuanto mayor sea la amplitud de su escala respecto a la otra. Sin embargo, planteando la ecuación en estandarizadas (véase cuadro de coeficientes):

$$\hat{Y} = 0.32X_1 + 0.721X_2$$

los coeficientes son directamente comparables. En este caso diremos que el efecto esperado (en desviaciones típicas) de la nota a cargo de la motivación (X_2) es de 0.401 (0.721-0.32) desviaciones típicas más que el cambio esperado por el efecto exclusivo de la variable inteligencia.

2.4. Validez del modelo.

El ANOVA y su correspondiente índice F es otra de las informaciones proporcionadas por el comando regresión en el SPSS para el modelo múltiple. La tabla de resultados obtenida para nuestros datos es:

ANOVA(b)

Modelo		Suma de cuadrados	Gl	Media cuadrática	F	Sig.
1	Regresión	36,238	2	18,119	25,368	,000(a)
	Residual	6,428	9	,714		
	Total	42,667	11			

a Variables predictoras: (Constante), motivación, Inteligencia
b Variable dependiente: nota

Como vemos, el modelo se muestra claramente válido para representar los datos. El valor de significación obtenido (p=.000) indica que la probabilidad de que el conjunto de variables predictoras introducidas no sea suficiente para aportar explicación de los valores predichos de Y es nula. Es decir, las variaciones en la variable nota se explican significativamente por el conjunto de efectos predictivos identificados.

Sin embargo, un análisis más pormenorizado sobre cada uno de estos predictores (como se tiene en la tabla anterior de coeficientes) nos conduce a afirmar que no todos ellos resultan igualmente relevantes. Si es así, quizás sea conveniente llevar a cabo una depuración del modelo de regresión múltiple eliminando aquellos regresores que por sus poderes predictivos –no significativos- no expliquen mucho más sobre Y que lo que ya explican aquellos que sí lo son. Veamos a continuación cómo puede llevarse a cabo esta tarea.

2.5. Depuración del modelo.

2.5.1. Método “introducir”.
Sabemos por los resultados obtenidos anteriormente (mediante el método “introducir”) que en nuestro modelo sólo la variable motivación ha resultado significativa en la ecuación para predecir la nota. Si en base a esta información eliminásemos del modelo la variable inteligencia y calculásemos la ecuación de regresión simple de la motivación sobre la nota tendríamos:

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	3,295	,595		5,536	,000
	Motivación	,133	,023	,878	5,814	,000

a Variable dependiente: nota

Unos datos que definen una ecuación del tipo:

$$\hat{Y} = 3.295 + .133X$$

lo que indica que en esta ecuación de regresión simple la nota se incrementa .133 puntos por cada unidad incrementada en la variable motivación (recuérdese que cuando dicha variable compartía la explicación de la nota junto con la inteligencia en la ecuación múltiple su coeficiente parcial era .109, menor que el de ahora).

Y la validez de dicho modelo se prueba en la siguiente tabla de ANOVA donde el valor de p asociado al índice F nos indica una significación de este coeficiente:

ANOVA(b)

Modelo		Suma de cuadrados	Gl	Media cuadrática	F	Sig.
1	Regresión	32,927	1	32,927	33,807	,000(a)
	Residual	9,740	10	,974		
	Total	42,667	11			

a Variables predictoras: (Constante), motivación
b Variable dependiente: nota

Efectivamente, el modelo resulta válido y por tanto la nota está afectada significativamente por la variable motivación.

2.5.2. Método por pasos.

Si hubiésemos calculado la ecuación de regresión múltiple mediante el método por pasos (stepwise) en vez de mediante el método "introducir", la salida obtenida nos hubiera informado desde el principio sobre la significación o no de cada variable predictora en el modelo calculando el poder de predicción del modelo una vez eliminada aquella variable no relevante en el mismo. Las instrucciones concretas para activar esta operación son: Analizar/Regresión /Lineal/ variable dependiente "nota" e independientes "inteligencia y motivación" especificando en la casilla de método de introducción el tipo "stepwise" (método por pasos):

2. Regresión lineal múltiple

Variables introducidas/eliminadas(a)

Modelo	Variables introducidas	Variables eliminadas	Método
1	Motivación	.	Por pasos (criterio: Prob. de F para entrar <= ,050, Prob. de F para salir >= ,100).

a Variable dependiente: nota

Se especifica que la única variable contemplada en la ecuación ha sido la motivación dado que únicamente su coeficiente satisface la condición de tener asociado un valor de p menor a .05. De esta forma, el coeficiente de correlación simple (dado que sólo existe dicha variable en la ecuación) con la nota es de .878 (el obtenido en el análisis de las correlaciones bivariadas) y su cuadrado .772, tal y como indica el siguiente cuadro:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,878(a)	,772	,749	,98690

a Variables predictoras: (Constante), motivación

El modelo además resulta significativo con una p=.000 (el mismo del ANOVA de antes calculado con la ecuación simple motivación-nota):

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	32,927	1	32,927	33,807	,000(a)
	Residual	9,740	10	,974		
	Total	42,667	11			

a Variables predictoras: (Constante), motivación
b Variable dependiente: nota

Los coeficientes de la ecuación de regresión simple se toman de la siguiente tabla:

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	3,295	,595		5,536	,000
	motivación	,133	,023	,878	5,814	,000

a Variable dependiente: nota

Nótese que la ecuación de regresión es, tal y como estimábamos antes a través del modelo de regresión simple al eliminar la inteligencia:

$$\hat{Y} = 3.295 + 0.133X$$

Por último, el siguiente cuadro proporciona información sobre la variable inteligencia excluida del modelo:

Variables excluidas(b)

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
						Tolerancia
1	Inteligencia	,320(a)	2,153	,060	,583	,757

a Variables predictoras en el modelo: (Constante), motivación
b Variable dependiente: nota

En primer lugar se informa sobre el coeficiente estandarizado de la inteligencia si dicha variable hubiera sido introducida en el modelo múltiple (.320) -el valor que ya sabíamos de antes-, su correspondiente valor t (2.153) y p asociado (.06), un valor mayor a .05, de ahí su no significación. En la siguiente casilla se informa sobre la correlación parcial de la inteligencia con la nota (.583) que se refiere a la correlación entre ambas cuando tanto de la inteligencia como de la nota se elimina –o no se tiene en cuenta- lo que comparten con la motivación. El valor indicado de correlación parcial (que ya calculamos al principio – véase apartado de coeficientes-) no es muy alto, lo esperado, dado que si lo hubiera sido la variable inteligencia hubiera sido incorporada al modelo como predictor relevante.

Finalmente, el índice de tolerancia del cuadro anterior se refiere al complementario de la correlación entre inteligencia y motivación (las dos variables predictoras del modelo) de tal manera que valores bajos de este estadístico representan correlaciones altas entre ellas y valores altos a la inversa. Como ya hemos tenido ocasión de probar, hay que tener en cuenta que cuanto mayor es la correlación entre las variables predictoras de un modelo mayormente quedan afectados sus coeficientes de correlación parcial correspondientes y por tanto más cambios sufren los coeficientes de regresión estimados en el modelo de regresión múltiple. El índice de tolerancia se cuantifica así:

$$1 - R_{12}^2$$

siendo R_{12}^2 el coeficiente de correlación al cuadrado entre las variables X_1 y X_2, es decir, entre inteligencia y motivación para nuestro caso:

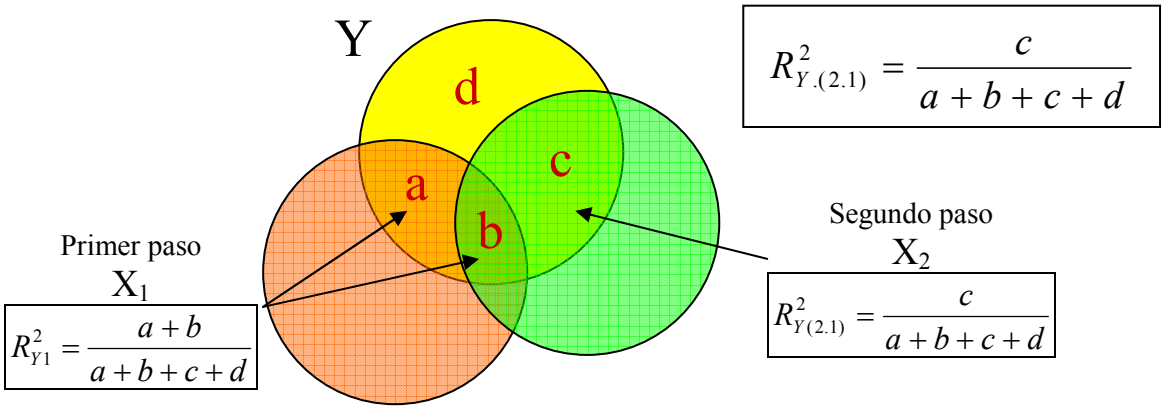
$$1 - 0.493^2 = 0.757$$

2.5.3. Correlación semiparcial.

Al hilo de la estrategia desarrollada mediante el método de estimación por pasos, es importante referirnos en este momento al concepto de la *correlación semiparcial* entre variables. Este cálculo permite conocer el poder explicativo de cada predictor en el modelo de regresión cuando éste se incorpora después de otro u otros regresores ya contemplados. De esta forma se decide si ampliar o no el número de variables explicativas al modelo en función de que los coeficientes de correlación semiparcial de éstas (incorporadas de nuevo) sean significativos o no lo sean.

Si representamos en un diagramo de Venn la medida de la correlación semiparcial al cuadrado asociado a una variable X_2 cuando ésta se incorpora como nueva a un modelo

en el que ya existe como predictor X_1 ($R^2_{Y(2.1)}$ - proporción de variabilidad explicada por X_2 sumada a lo que ya explicaba X_1-) tendremos:



En este diagrama puede observarse que si la parte explicada de Y por parte de X_1 es (a + b), lo que suma la incorporación de la X_2 a la explicación de Y es precisamente el área de Y que no explicaba antes X_1 (c). Así pues, en términos de áreas tendremos:

$$R^2_{Y(2.1)} = \frac{a+b+c}{a+b+c+d} - \frac{a+b}{a+b+c+d} = \frac{c}{a+b+c+d}$$

Justamente lo que aporta de nuevo la variable X_2 a la explicación de Y.

Formalmente el coeficiente de correlación semiparcial al cuadrado se expresa así:

$$R^2_{Y.(2.1)} = R^2_{Y.12} - R^2_{Y.1}$$

esto es, la porción de variabilidad explicada por X_2 de Y (de todo Y) una vez eliminada la que explicaba ya X_1.

Para nuestros datos, calculemos el coeficiente de correlación semiparcial de la variable nota con la inteligencia (X_1), suponiendo que ésta se introduce como mueva en un modelo en el que ya existía la motivación (X_2). Se sabe que la porción de variabilidad explicada de Y por la motivación (X_2) es .772 y el coeficiente de correlación múltiple al cuadrado (con las dos variables introducidas) es .849: Así:

$$R^2_{Y.2} = .772$$
$$R^2_{Y.12} = .849$$

Por lo que el incremento en la variabilidad explicada de Y al introducir como nueva la variable X_1 (inteligencia) es:

$$R^2_{Y(1.2)} = R^2_{Y.12} - R^2_{Y.2} = .849 - .772 = .077$$

Este valor (.077) es justamente el coeficiente de correlación semiparcial al cuadrado de la variable inteligencia con la nota; esto es, lo que añade de explicación esta variable

sobre lo que ya aportaba la motivación. Si obtenemos la raíz cuadrada de este valor obtenemos el coeficiente de correlación semiparcial de inteligencia con la nota:

$$r_{Y(2.1)} = \sqrt{.079} = .277$$

que indica su correlación con las notas una vez eliminada aquella correlación que éstas mantienen con la motivación.

Sabemos por los análisis realizados antes que este incremento en variabilidad explicada por parte de la inteligencia sobre la motivación no resulta estadísticamente significativa de ahí que se considerara aconsejable eliminar dicha variable del modelo.

Veamos, sin embargo, cómo se valida exactamente la significación de dicho coeficiente de correlación semiparcial. Puede utilizarse una F para satisfacer este objetivo, una F que evalúa la significación de este incremento en la explicación de Y:

$$F = \frac{\dfrac{R^2_{Y.12} - R^2_{Y(2.1)}}{K_1 - K}}{\dfrac{1 - R^2_{Y.12}}{N - K - 1}}$$

Los grados de libertad asociados a esta F son: los de la diferencia entre el número total de variables contempladas en el modelo y aquélla o aquéllas incorporadas de nuevo (para numerador) y el número total de variables (en el denominador).

En nuestro caso tendríamos:

$$F = \frac{\dfrac{.849 - .772}{2 - 1}}{\dfrac{1 - .849}{12 - 2 - 1}} = \frac{.077}{.0167} = 4.61$$

Comparando esta F empírica con su correspondiente teórica para 1 y 9 grados de libertad y α=.05 que es igual a 5.12, nos conduce a aceptar la hipótesis nula (.509<5.12) y concluir que la correlación semiparcial de la inteligencia con la nota no es estadísticamente significativa. Es decir, la incorporación de la variable inteligencia al modelo en el que ya se contemplaba la motivación no aumenta significativamente la explicación de Y.

En el SPSS podemos solicitar el cálculo de la correlación semiparcial de cualquier variable con la Y marcando esta opción dentro del apartado de "estadísticos". Si así lo hacemos, la salida correspondiente para nuestros datos sería la siguiente:

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coef. Estandarizados	t	Sig.	Correlaciones		
		B	Error típ.	Beta			Orden cero	Parcial	Semiparcial
1	(Constante)	1,737	,885		1,964	,081			
	Inteligencia	,019	,009	,320	2,153	,060	,675	,583	,279
	motivación	,109	,022	,721	4,847	,001	,878	,850	,627

a Variable dependiente: nota

Podemos apreciar que el coeficiente de correlación semiparcial apuntado para la inteligencia coincide prácticamente con el calculado arriba (.279 ≅.277) y su p=.060 implica la no significación de dicho coeficiente (.060>.05), soluciones que coinciden con lo afirmado entonces.

Por otro lado, el valor del incremento en R^2 asociado a cada variable del modelo puede también solicitarse como resultado al SPSS. Para ello es necesario que ambas variables resulten significativas en el modelo, algo que no ocurre en nuestro caso. Sin embargo, aumentando el valor de la significación permitida para ellas (desde .05 a .07) logramos que nos muestre dicho incremento asociado a cada una de ellas resultando la siguiente salida:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,878(a)	,772	,749	,98690	,772	33,807	1	10	,000
2	,922(b)	,849	,816	,84514	,078	4,636	1	9	,060

a Variables predictoras: (Constante), motivación
b Variables predictoras: (Constante), motivación, Inteligencia

En el modelo 1 se especifican los estadísticos referidos exclusivamente a la variable motivación por ser la más significativa en el modelo múltiple. Como puede apreciarse, la R^2, su F y significación coinciden con los calculados previamente cuando solamente existía esta variable en el modelo. ¿Cuáles son los cambios previstos si se introduce además de ésta la variable inteligencia para explicar Y –modelo 2-?. Pues que el valor de R^2 aumenta hasta .849, la diferencia respecto a la R^2 anterior es de .078 (tal y como sabíamos de antes), la F de este incremento es 4.636 (el valor calculado antes era 4.61) y la significación asociada a la misma de .06 (de ahí la no inclusión de esta variable en el modelo pues supera el valor de significación permitido -.05-)

Obsérvese que las significaciones asociadas a la variable inteligencia, tanto respecto a su correlación parcial como semiparcial coinciden (.06). Son, pues, dos caminos para averiguar si a un determinado nivel de significación resulta relevante introducir una variable en el modelo.

Habiendo desarrollado en estas páginas los conceptos de correlación parcial y semiparcial entre variables puede decirse, en resumen, que mientras la correlación semiparcial de una variable con la criterio mide su contribución particular –después de lo explicado por otras- a la explicación de dicho criterio (en su conjunto) , la correlación parcial se refiere a su correlación con el criterio una vez eliminado del mismo todo lo que tiene que ver con las restantes variables consideradas en el modelo. Es decir, su correlación con lo que queda del criterio independiente a la información que comparte éste con las restantes.