Técnicas de regresión: Regresión Lineal Múltiple

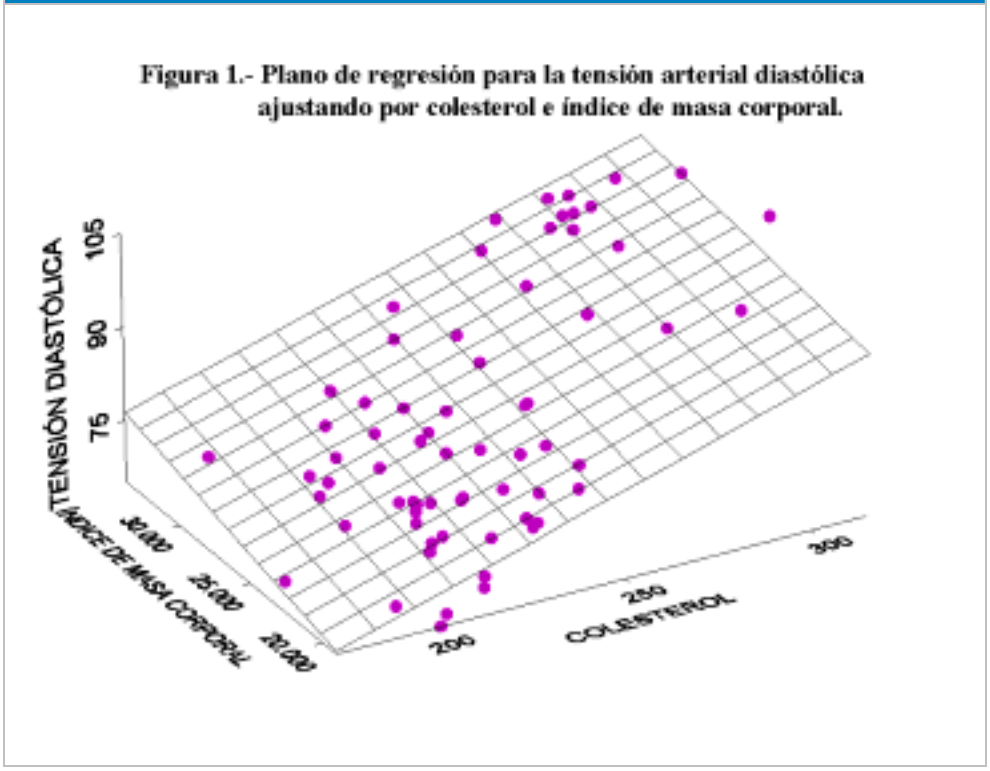
Pértega Díaz S., Pita Fernández S.
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña.
Cad Aten Primaria 2000; 7: 173-176.
Actualización 20/08/2001.

La mayoría de los estudios clínicos conllevan la obtención de datos en un número más o menos extenso de variables. En algunos casos el análisis de dicha información se lleva a cabo centrando la atención en pequeños subconjuntos de las variables recogidas utilizando para ello análisis sencillos que involucran únicamente técnicas bivariadas. Un análisis apropiado, sin embargo, debe tener en consideración toda la información recogida o de interés para el clínico y requiere de técnicas estadísticas multivariantes más complejas. En particular, hemos visto como el modelo de regresión lineal simple es un método sencillo para analizar la relación lineal entre dos variables cuantitativas. Sin embargo, en la mayoría de los casos lo que se pretende es predecir una respuesta en función de un conjunto más amplio de variables, siendo necesario considerar el modelo de regresión lineal múltiple como una extensión de la recta de regresión que permite la inclusión de un número mayor de variables.

Estimación de parámetros y bondad de ajuste.

Generalizando la notación usada para el modelo de regresión lineal simple, disponemos en n individuos de los datos $\{(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)\}_{i=1,\ldots,n}$ de una variable respuesta Y y de p variables explicativas $X_1, X_2, \ldots, X_p$. La situación más sencilla que extiende el caso de una única variable regresora es aquella en la que se dispone de información en dos variables adicionales. Como ejemplo, tomemos la medida de la tensión arterial diastólica en setenta individuos de los que se conoce además su edad, colesterol e índice de masa corporal (Tabla 1). Es bien conocido que el valor de la tensión arterial diastólica varía en función del colesterol e índice de masa corporal de cada sujeto. Al igual que ocurría en el caso bidimensional, se puede visualizar la relación entre las tres variables en un gráfico de dispersión, de modo que la técnica de regresión lineal múltiple proporcionaría el plano que mejor ajusta a la nube de puntos resultante (Fig. 1).

Figura 1. Plano de regresión para la Tensión Arterial Diastólica ajuntando por Colesterol e Índice de Masa Corporal



Del gráfico se deduce fácilmente que los pacientes con tensión arterial diastólica más alta son aquellos con valores mayores de colesterol e índice de masa corporal. Si el número de variables explicativas aumenta (p>2) la representación gráfica ya no es factible, pero el resultado de la regresión se generaliza al caso del mejor hiperplano que ajusta a los datos en el espacio (p+1)-dimensional correspondiente.

Tabla 1. Edad, Colesterol, Índice de Masa Corporal y Tensión Arterial Diastólica de 70 pacientes.

	EDAD	COLESTEROL	IMC	TAD		EDAD	COLESTEROL	IMC	TAD
1	42	292	31,64	97	36	53	187	23,31	80
2	64	235	30,80	90	37	43	208	27,15	65
3	47	200	25,61	80	38	57	246	21,09	80
4	56	200	26,17	75	39	64	275	22,53	95
5	54	300	31,96	100	40	43	218	19,83	75
6	48	215	23,18	67	41	47	231	26,17	75
7	57	216	21,19	,	42	58	200	25,95	90
8	52	254	26,95	70	43	58	214	26,30	75
9	67	310	24,26	105	44	48	230	24,89	70
10	46	237	21,87	70	45	62	280	26,89	100
11	58	220	25,61	70	46	54	198	21,09	65
12	62	233	27,92	75	47	67	285	31,11	95
13	49	240	27,73	90	48	68	201	21,60	80
14	56	295	22,49	95	49	55	206	19,78	65
15	63	310	,	95	50	50	223	22,99	75
16	64	268	30,04	90	51	53	290	32,32	95
17	67	243	23,88	85	52	63	315	31,14	100
18	49	239	21,99	75	53	60	220	28,89	80
19	53	198	26,93	75	54	46	230	20,55	75
20	59	218	,	85	55	45	175	22,49	70
21	65	215	24,09	70	56	53	213	22,53	70
22	67	254	28,65	105	57	59	220	20,82	65
23	49	218	25,71	85	58	62	287	32,32	95
24	53	221	25,33	80	59	60	290	33,91	90
25	57	237	25,42	90	60	62	209	20,76	75
26	47	244	23,99	85	61	58	290	31,35	80
27	58	223	25,20	70	62	57	260	31,14	95
28	48	198	25,81	85	63	49	202	20,76	80
29	51	234	26,93	80	64	61	214	19,59	90
30	49	175	27,77	80	65	52	231	20,08	75
31	68	230	30,85	70	66	59	280	31,60	100
32	58	248	21,61	75	67	50	220	25,34	70
33	54	218	26,30	95	68	46	233	22,86	75
34	59	285	31,44	100	69	44	215	19,53	70
35	45	253	25,00	75	70	60	202	19,10	65

En el caso general, el modelo de regresión lineal múltiple con p variables responde a la ecuación:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \quad i = 1,\ldots,n \qquad (1)$$

de modo que los coeficientes β_i se estiman siguiendo el criterio de mínimos cuadrados:

$$\min_{\substack{\beta_i \in \mathbb{R} \\ i=1,\dots,n}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \dots - \beta_p X_{ip})^2$$

La obtención aquí de las expresiones de los estimadores mínimo cuadráticos de dichos coeficientes exigen reescribir la expresión (1) utilizando notación matricial. Así, (1) quedaría:

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

donde:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2P} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{(n-1)1} & X_{(n-1)2} & \dots & X_{(n-1)p} \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \quad y \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

De donde los estimadores mínimo cuadráticos se obtienen a partir de la ecuación:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

y mantienen una interpretación análoga al caso de la regresión lineal simple (i.e. $\hat{\beta}_i$ representa el incremento por término medio en la variable respuesta por cada unidad adicional en la variable X_i). Como se puede observar, la obtención de estimadores, intervalos de confianza y contrastes de hipótesis para los coeficientes de regresión involucran expresiones matriciales y distribuciones multivariantes que complican notablemente las operaciones, por lo que en la práctica dichos cálculos se obtienen de un modo inmediato mediante el manejo de diferentes paquetes estadísticos. Son muchos los textos en los que se pueden encontrar desarrollos teóricos de dichas expresiones[1],[2]. Sin detenerse en ello, basta decir que manteniendo las hipótesis habituales de independencia, homocedasticidad, normalidad y linealidad se calculan expresiones para el error estándar de cada coeficiente estimado e intervalos de confianza de modo análogo al caso de la regresión simple. La significación estadística de cada variable se obtiene simplemente calculando el cociente entre el coeficiente estimado y su error típico, y comparándolo con el cuantil correspondiente de una distribución t de Student con n-p-1 grados de libertad. La bondad de ajuste del modelo se puede valorar mediante la varianza residual y el estadístico R^2 (*coeficiente de determinación*), definidos de la forma habitual. También aquí puede utilizarse el *contraste F global de la regresión*, calculado a partir de las sumas de cuadrados explicada y no explicada para valorar la utilidad del modelo. Como ejemplo, tras ajustar un modelo de regresión múltiple a los datos que se muestran en la Tabla 1 usando como variables predictoras de la tensión diastólica el colesterol e índice de masa corporal de un individuo, los coeficientes de regresión para ambas variables fueron 0.18 (E.T. 0.03) y 0.73 (E.T. 0.30) respectivamente, siendo ambos significativamente distintos de cero (Tabla 2). Esto indica que por término medio la tensión arterial diastólica de un paciente se incrementa en 1.8 y 7.3 respectivamente por cada 10 unidades a mayores en su colesterol o índice de masa corporal. El valor del coeficiente de determinación R^2=52% y la significación del contraste F global de la regresión ($p<0.001$) sugieren que gran parte de la variabilidad de la respuesta viene explicada por el modelo ajustado.

Tabla 2. Modelo de regresión lineal múltiple para la tensión arterial diastólica ajustando por colesterol e índice de masa corporal.					
Variable	**Coeficiente (B)**	**E.T.(B)**	**IC 95% (B)**	**t**	**p**
Constante	19.42	7.54	(4.37;34.48)	2.58	0.012
Colesterol	0.18	0.03	(0.11;0.25)	5.26	<0.001
IMC	0.73	0.30	(0.14;1.33)	2.45	0.017
	Suma de Cuadrados	**g.l.**	**Media cuadrática**	**F**	**p**
Regresión	4,449.72	2	2,224.86	34.93	<0.001
Residual	4,076.40	64	63.69		
Total	8,526.12	66			

El hecho de contar con un número más extenso de variables exige que además del contraste F global se puedan realizar pruebas parciales para constatar si un grupo de variables añadidas a un modelo lo mejoran. Supongamos que al modelo (1) se suma una nueva variable explicativa X^*. La proporción de variabilidad residual que es explicada al introducir esta nueva variable viene dada por la diferencia en las sumas de cuadrados de cada modelo:

$$SC\,\mathrm{Re}\,gresión(X^* / X_1,\ldots,X_p) = SC\,\mathrm{Re}\,gresión(X_1,\ldots,X_p,X^*) - SC\,\mathrm{Re}\,gresión(X_1,\ldots,X_p)$$

Para valorar si la introducción de la nueva variable queda compensada por una mejora significativa en la predicción de la respuesta se utiliza el estadístico:

$$F = \frac{SC\,\mathrm{Re}\,gresión(X^* / X_1,\ldots,X_p)}{\dfrac{SC\,\mathrm{Re}\,sidual(X_1,\ldots,X_p,X^*)}{n-p-2}}$$

que se compara con el cuantil correspondiente de una distribución F de Snedecor con 1 y n-p-2 grados de libertad. Dicho contraste se denomina *contraste F parcial*. Para comprobar el uso de dicho estadístico consideremos en el ejemplo anterior el modelo de regresión simple que resulta de tomar como única variable regresora el colesterol de un individuo (Tabla 3). El valor del estadístico R^2 en este caso es del 69.1% frente al 72.2% del modelo que se consigue introduciendo el índice de masa corporal como nueva variable explicativa. El cambio en el estadístico R^2 es de 0.045 que coincide con el cuadrado del coeficiente de correlación parcial entre la tensión arterial y el índice de masa corporal ajustando por el colesterol. La significación del contraste F parcial para la introducción del índice de masa corporal es de 0.017, indicando que el modelo con dos variables mejora al modelo más simple.

Tabla 3. Modelo de regresión lineal simple para la tensión arterial diastólica ajustando por colesterol.					
Variable	**Coeficiente (B)**	**E.T.(B)**	**IC 95% (B)**	**t**	**p**
Constante	26.91	7.15	(12.63;41.19)	3.76	<0.001
Colesterol	0.23	0.03	(0.17;0.29)	7.70	<0.001
	Suma de Cuadrados	**g.l.**	**Media cuadrática**	**F**	**p**
Regresión	4,067.11	1	4,067.11	59.29	<0.001
Residual	4,459.01	65	68.60		
Total	8,526.12	66			

Es importante recalcar la necesidad de uso de métodos estadísticos multivariantes para estudiar correctamente la relación entre más de dos variables. La aplicación de las técnicas de regresión ha sido tratada en diversos textos[3],[4],[5],[6] desde un punto de vista eminentemente práctico. Aunque el modelo de regresión se ha planteado inicialmente para analizar la relación entre variables cuantitativas, su generalización al caso de variables regresoras cualitativas es inmediata. Este tipo de análisis recibe el nombre de análisis de covarianza o análisis de varianza según contenga o no además variables numéricas. La limitación de este modelo por considerar que la relación de cada variable con la respuesta es de tipo

lineal queda solventada mediante la transformación (logarítmica, cuadrática,...) de cada variable regresora.

Selección de variables.

Una de las principales dificultades a la hora de ajustar un modelo de regresión múltiple surge cuando es necesario identificar entre el conjunto de variables disponibles aquellas que están relacionadas con la respuesta y que la predicen de la mejor forma posible. Cuando el número de variables es reducido, como en el ejemplo manejado, la selección no resulta complicada. Una primera alternativa es construir un modelo por inclusión o hacia delante ("forward"), considerando en primer lugar la relación de cada variable con la respuesta e ignorando todas las demás variables, valorándola por medio del coeficiente de correlación lineal de Pearson (Figura 2). Aquella que muestra una correlación más alta con la variable dependiente (en este caso el colesterol) se introduce en un modelo inicial (Tabla 3).

El segundo paso consiste en seleccionar entre las variables restantes aquella que al introducirla en el modelo permite explicar una mayor parte de la variabilidad residual. La comparación entre distintos modelos debe hacerse en términos del valor relativo de los coeficientes de determinación y el contraste F parcial. Ya vimos como la inclusión del índice de masa corporal reportaba una mejora en el modelo de regresión simple. La introducción de la variable edad, en cambio, proporciona un cambio en el coeficiente de determinación de 0.028 que no resulta en una mejora significativa (p=0.059). Este esquema se repetiría hasta que ninguna otra variable entrase a formar parte del modelo. En el ejemplo el último paso sería comprobar si la introducción de la variable edad produce una mejora del ajuste del modelo mostrado en la Tabla 2. El modelo ajustando por las tres variables se muestra en la Tabla 4. El coeficiente correspondiente a esta última variable no es significativo (nótese que esta significación ha de coincidir con la del contraste F parcial correspondiente).

Tabla 4. Modelo de regresión lineal múltiple para la tensión arterial diastólica ajustando por colesterol, índice de masa corporal y edad.

Variable	Coeficiente (B)	E.T.(B)	IC 95% (B)	t	p
Constante	10.55	9.13	(-7.70;28.81)	1.15	0.252
Colesterol	0.17	0.03	(0.1;0.24)	4.84	<0.001
IMC	0.68	0.30	(0.09;1.28)	2.31	0.024
Edad	0.24	0.14	(-0.05;0.53)	1.67	0.100
	Suma de Cuadrados	**g.l.**	**Media cuadrática**	**F**	**p**
Regresión	4,622.52	3	1,540.84	24.87	<0.001
Residual	3,903.60	63	61.96		
Total	8,526.12	66			

En la mayoría de los casos se dispone de información en un conjunto mucho más amplio de variables de las que se desconoce cuáles están relacionadas o pueden utilizarse para predecir la respuesta de interés. La identificación del conjunto de variables que proporcionan el mejor modelo de regresión dependerá en gran medida del objetivo del estudio y de experiencias previas. Así, aunque la práctica habitual es eliminar del modelo aquellas variables que no resultan significativas, puede ser recomendable mantenerlas en caso de que en experiencias previas se haya constatado una relación con la variable dependiente. La mayoría de paquetes estadísticos proporcionan una variedad de técnicas para identificar el mejor conjunto de variables regresoras que introducen o eliminan sucesivamente variables atendiendo a su significación en el modelo (hacia delante, hacia atrás, pasos sucesivos). Existen otras alternativas basadas en la comparación de todos los modelos posibles que se pueden formar con un conjunto inicial de variables. Todas estas técnicas deben considerarse meramente orientativas. Así, identificado el mejor conjunto de variables y ajustado el modelo es conveniente realizar un análisis de residuos exhaustivo para valorar la posibilidad de elegir un modelo distinto a pesar de que tenga un valor menor de R^2.

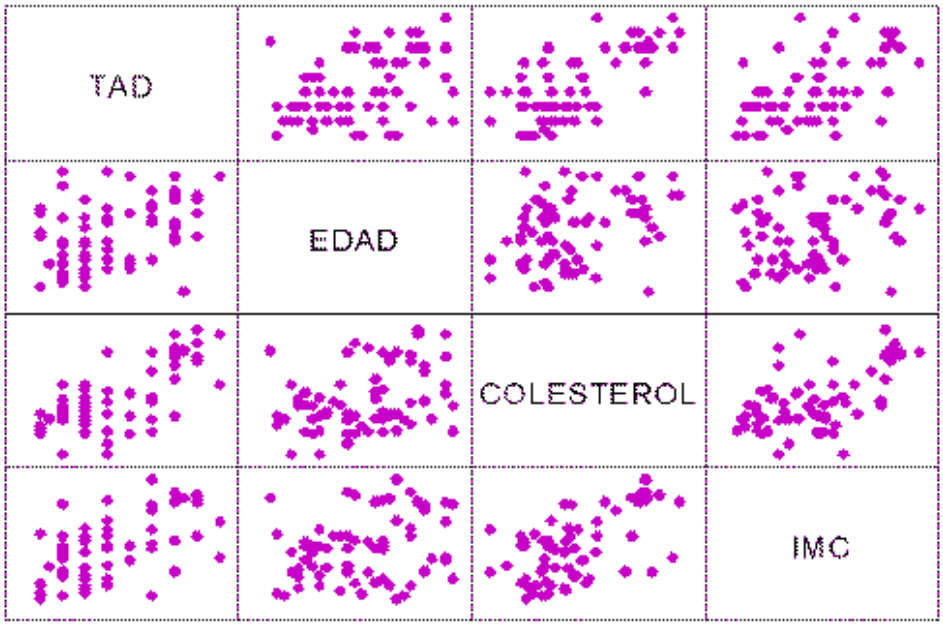
Interacción, confusión y colinealidad.

Cuando se introduce más de una variable en el modelo de regresión es necesario contrastar además la independencia de los efectos de todas ellas. Es decir, se supone que la asociación de cada variable con la respuesta no depende del valor que tomen el resto en la ecuación de regresión. En otro caso se dirá que

existe *interacción.* Antes de aprobar el modelo definitivo, por lo tanto, se debe explorar la necesidad de incluir términos de interacción calculados a partir del producto de pares de variables, comprobando si mejora la predicción, siendo aconsejable investigar solamente aquellas interacciones que puedan tener una explicación clínica.

En ocasiones el fenómeno de la interacción se hace coincidir erróneamente con los de *confusión* y *correlación*. Existe *confusión* cuando el efecto de una variable difiere significativamente según se considere o no en el modelo alguna otra. Ésta se asociará tanto con la variable inicial como con la respuesta, de modo que en casos extremos puede invertir el primer efecto observado. En ese caso las estimaciones adecuadas son aquellas que proporciona el modelo completo, y se dirán que están controladas o ajustadas por variables de confusión. Por otro lado, el fenómeno que se produce cuando dos variables explicativas muestran una correlación alta recibe el nombre de cuasi-colinealidad y puede producir estimaciones inestables de los coeficientes que se traducen en valores desorbitados de sus errores típicos y resultados poco creíbles. La mayoría de paquetes estadísticos muestran en sus salidas diagnósticos de colinealidad (tolerancia, factor de inflacción de la varianza, índice de condición) que pueden ayudarnos a solventar estos problemas. Por lo tanto, se ha de tener un cuidado especial en la etapa de construcción del modelo: un cambio significativo en las estimaciones tras la inclusión de una nueva variable puede evidenciar cualquiera de estos fenómenos. Nos corresponde a nosotros evaluar la conveniencia de incluirla o no en el modelo.

Figura 2. Correlación lineal de la edad, colesterol e índice de masa corporal con la tensión arterial diastólica.



Bibliografía

1. Snedecor G.W., Cochran W.G. Statistical Methods. 8th ed. Iowa State University Press; 1989.
2. Seber GAF. Linear Regression Analysis. New York: John Wiley & Sons; 1977.
3. Etxebarría Murgiondo, J. Regresión Múltiple. Madrid: La Muralla; 1999.
4. Altman DA. Practical statistics for medical research. 1th ed., repr. 1997. London: Chapman & Hall; 1997.
5. Carrasco J.L., Hernán M.A. Estadística Multivariante en las Ciencias de la Salud. Madrid: Ed. Ciencia 3; 1993.
6. Kleinbaum D.G., Kupper L.L. Applied Regression Analysis and other Multivariable Methods. 3rd. ed. Massachusetts: Duxbury Press; 1997.