Rincón de la Bioestadística

Modelación estadística: La regresión lineal simple. Parte 2

Gabriel Cavada Ch.^{1,2}

¹División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.

²Facultad de Medicina, Universidad de los Andes.

Statical modeling: Simple linear regression

Siguiendo la idea inspiradora de la regresión lineal simple, expuesta en el número anterior de esta revista, esto es que al observar un efecto y buscar la o las causas que lo produjeron, aceptamos ampliamente que este efecto puede ser producido por más de una causa, es decir, aceptamos la multicausalidad de una respuesta, concepto que nos acerca con mayor fidelidad a la realidad que estamos observando. Si la respuesta o efecto lo atribuimos a “p” posibles causas, la idea anterior la expresamos matemáticamente así:

$$Y = f(X_1, X_2, X_3, \dots, X_p)$$

Nos interesa modelar la respuesta cuando la relación funcional entre la respuesta y la causa es lineal, es decir, de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_p X_p$$

o bien, si nos centramos en la i-ésima observación:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_p X_{pi} + \varepsilon_i$$

La estimación de los parámetros del modelo se obtienen por el método de máxima verosimilitud: Se basa en suponer que:

$$Y_i \sim N(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_p X_{pi}, \sigma^2)$$

o equivalentemente:

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

Así la función de verosimilitud toma la siguiente forma:

$$L(\beta_0, \beta_1, \dots, \beta_p, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma}} \right)^n e^{-\frac{1}{2\sigma^2} \sum \left(\frac{Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_p X_{pi})}{\sigma} \right)^2}$$

Al igual que en regresión lineal simple, se obtienen los estimadores de los parámetros del modelo. Sin embargo, es necesario agregar el supuesto de independencia entre los predictores; esta exigencia se conoce como “no colinealidad” o no multicolinealidad. De existir este problema, una variable que es en sí misma explicativa de la respuesta puede aparecer como no significativa ya que el efecto de alta colinealidad se refleja en una estimación con mucha varianza y por lo tanto con intervalos de confianza más anchos de lo que en la realidad son.

Una vez ajustado un modelo de regresión, es necesario conocer la calidad del mismo; para ello la variabilidad total de Y, que no depende del modelo ajustado, puede descomponerse del siguiente modo:

$$\sum (Y - \bar{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2$$

$$SCTotal = SCResidual + SCRegresión$$

$$\text{Varianza Total} = \text{Varianza no explicada} + \text{Varianza explicada}$$

Se define el coeficiente de determinación como:

$$R^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$

Como la respuesta está siendo explicada por más de una variable predictora, no es posible definir un coeficiente de correlación como el conocido “r” de Fisher, usado en regresión lineal simple. Por ello estudiamos con detalle el coeficiente de determinación R². Sin embargo, el coeficiente de determinación, aumenta en la medida que se incorporan variables al modelo, sin aportar, necesariamente una explicación plausible a la respuesta; por esta razón es necesario corregir este coeficiente, por la cantidad de predictores involucrados en el modelo; a este coeficiente de determinación corregido se le llama R² ajustado.

La descomposición de la variabilidad o Tabla ANOVA es: (Tabla 1).

Rincón de la Bioestadística

Tabla 1.

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Regresión	p	$\sum(\hat{Y} - \bar{Y})^2$	$CMreg = \frac{\sum(\hat{Y} - \bar{Y})^2}{p}$	$F = \frac{CMreg}{CMres}$
Residuo	n-p-1	$\sum(Y - \hat{Y})^2$	$CMres = \frac{\sum(Y - \hat{Y})^2}{n - p - 1}$	
Total	n-1	$\sum(Y - \bar{Y})^2$		

Asociada a la descomposición de la variabilidad y por ende a la calidad del modelo, se tiene la siguiente dócima:

$$H_0 : \mathcal{R}^2 = 0$$

$$H_1 : \mathcal{R}^2 > 0$$

Es decir la hipótesis nula dice que el coeficiente de determinación es cero o equivalentemente el modelo de regresión lineal no existe, *versus* la hipótesis alternativa expresando que el coeficiente de determinación es positivo o decir equivalentemente que el modelo existe.

Cuya estadística de prueba es:

$$F = \frac{CMreg}{CMres} \sim F(p, n-p-1)$$

La F (p, n-p-1) indica el valor crítico en la distribución de Fisher-Snedecor.

Si la hipótesis nula es rechazada debemos concluir que a lo menos uno de los predictores involucrados en el modelo es significativo, en ningún caso significa que todos lo son.

La matemática involucrada en las estimaciones de los parámetros de la regresión múltiple es de mediana complejidad y no será detallada aquí.

Sin embargo, presentaremos un ejemplo ilustrativo que

muestra una estrategia de análisis mínima usando la regresión lineal múltiple:

Ejemplo ilustrativo:

En pacientes con antecedentes con infarto agudo al miocardio, se desea explicar la presión arterial media (pam=2) • (presión diastólica)/3-(presión sistólica)/3) a través de las siguientes variables:

Variable	Etiqueta
Edad	en años
IMC	Índice de masa corporal
Glicma	Glicemia
ColTot	Colesterol Total
HDL	Colesterol Hdl
LDL	Colesterol LDL
Trigl	Triglicéridos

Estrategia de análisis

1. Descripción de la muestra

Se analizan 3.296 pacientes de los cuales 2.422 (73,48%) son varones y 874 (26,52%) son mujeres.

La Tabla 2 muestra predictores continuos y la respuesta, descritos por: el mínimo, los cuartiles, el máximo, el promedio y la desviación estándar.

Tabla 2.

Variable	n	Min	p25	p50	p75	Max	Promedio	Desviación estándar
Edad	3.296	17	53	62	72	98	62,5	12,7
Glicma	3.296	31	112	134	183	926	162,8	84,2
ColTot	3.296	66	164	194	224	992	197,7	58,3
HDL	3.296	10,2	35	42	50	99	43,3	12,4
LDL	3.296	10	92	119	146	415	120,4	41,4
Triglic	3.296	18	98,5	136	192	3.900	159,5	118,6
IMC	3.296	18	24,4	26,6	29,3	44,5	27,1	3,9
Pam	3.296	38	87,6	10,15	115,3	203,3	10,1	21,1

Rincón de la Bioestadística

Exploración de asociaciones de la Pam:

- I. Como la variable sexo es dicotómica, la asociación con Pam se realiza mediante el test t de Student, encontrándose que las mujeres tienen una Pam significativamente más baja que los hombres ($p = 0,0002$)
- II. La asociación de la respuesta y los predictores continuos se logra observando una “Matriz de Correlaciones”, que muestra la respectiva correlación y el valor p que señala si la respectiva correlación es significativamente distinta de 0.

	Pam	Edad	Glicma	Col Tot	HDL	LDL	Trigl
Pam	1.0000						
Edad	-0.0610	1.0000					
	0.0005						
Glicma	-0.0740	0.0991	1.0000				
	0.0000	0.0000					
Col Tot	0.1125	-0.1504	-0.0068	1.0000			
	0.0000	0.0000	0.6963				
HDL	0.0546	0.1451	-0.0383	0.1116	1.0000		
	0.0017	0.0000	0.0280	0.0000			
LDL	0.1235	-0.1537	-0.0370	0.7388	0.0369	1.0000	
	0.0000	0.0000	0.0338	0.0000	0.0340		
Trigl	0.0701	-0.1970	0.0091	0.2783	-0.1941	0.0769	1.0000
	0.0001	0.0000	0.6034	0.0000	0.0000	0.0000	
IMC	0.1295	-0.1827	0.0637	0.0838	-0.0960	0.0656	0.1206
	0.0000	0.0000	0.0003	0.0000	0.0000	0.0002	0.0000

Los valores p en negritas indican que la Pam está correlacionada significativamente con todas las variables predictoras, pero además se observa que entre los predictores hay una fuerte evidencia de multicolinealidad.

- III. Una forma de explorar el nivel de multicolinealidad es mediante el método de los valores y vectores propios de la matriz compuesta sólo por los predictores; así, se genera un índice, $L = \frac{\lambda_{\max}}{\lambda_{\min}}$, que representa el cociente entre el valor mayor y el valor menor propio de la matriz de correlaciones construida con sólo los predictores, y este se compara según el siguiente criterio:

- Si $L < 30$ hay poca colinealidad
- Si $30 \leq L \leq 100$ colinealidad moderada
- Si $L > 100$ hay colinealidad fuerte

El detalle de este procedimiento es complejo y requiere que el investigador recurra a un bioestadístico.

Para el ejemplo estudiado se encuentra como cociente de valores propios 8,8, lo que indica escasa colinealidad; esto permite estimar una regresión con todos los predictores continuos (sin incluir el género aún).

Al estimar el modelo se obtiene la siguiente salida computacional de STATA 10,1, que interpretamos a continuación:

Source	SS	df	MS	Number of obs = 3.296		
				F (7, 3288)	=	21.34
Modelo	63511.2566	7	9073.03666	Prob > F	=	0.0000
Residual	1398179.44	3288	425.237055	R-squared	=	0.0435
				Adj R-squared	=	0.0414
Total	1461690.69	3295	443.608708	Root MSE	=	20.621
	Coef.	SE	t	P > t	[95% Conf. Intervalo]	
Pam						
Edad	-.023115	.0299411	-0.77	0.440	-.0818201	.0355901
Glicma	-.0186019	.0043097	-4.32	0.000	-.0270519	-.0101518
Col Tot	-.0014451	.009902	-0.15	0.884	-.0208598	.0179696
HDL	.1260887	.0306547	4.11	0.000	.0659845	.186193
LDL	.0539938	.0133053	4.06	0.000	.0279062	.0800813
Trigl	.0107094	.003369	3.18	0.001	.004104	.0173149
IMC	.6686379	.0937813	7.13	0.000	.4847621	.8525136
Cons	75.09872	3.863546	19.44	0.000	67.52352	82.67392

Evidencia de la existencia del modelo: a lo menos una de las variables explica a la pam. P-value de la décima para la existencia del modelo

Rincón de la Bioestadística

Los valores p en negritas indican que las respectivas variables explican significativamente a la Pam. La hipótesis nula es que el coeficiente respectivo es nulo.

Un modelo de regresión múltiple debe ser “parsimonioso”, es decir debe explicar lo máximo posible con un mínimo de variables predictoras.

En este análisis se observa que ni la edad ni el colesterol total explican significativamente a la Pam: Si son sacadas del modelo ¿Se pierde capacidad de explicación? Esta respuesta la entrega el Test de Razón de Verosimilitudes (Likelihood ratio test), cuya hipótesis nula es que el modelo inicial y el modelo reducido tienen la misma capacidad de explicación:

Al realizar el Test de Razón de Verosimilitudes, cuya hipótesis nula expresa que el modelo completo (con todas las variables) tiene igual capacidad de explicación que el modelo reducido (al que se le han sacado las variables que no aparecieron significativas); determina un valor $p = 0,7353$ indicando que el modelo reducido explica igual cantidad de varianza que el inicial.

Por lo tanto, el modelo final es:

Pam	Coef.	SE.	t	P> t	95%	Conf. Intervalo
Glicma	-.0190041	.0042777	-4.44	0.000	-.0273913	-.0106169
HDL	.1225339	.0297766	4.12	0.000	.0641514	.1809165
LDL	.0534757	.0087438	6.12	0.000	.0363319	.0706195
Trigl	.0108991	.0031131	3.50	0.000	.0047952	.0170029
IMC	.6795968	.0925668	7.34	0.000	.4981025	.8610911
Cons	73.32228	3.121049	23.49	0.000	67.20288	79.44167

Es preciso notar que todas las variables son significativas, sin embargo, el impacto sobre la Pam ha variado respecto del modelo inicial.

Para interpretar un coeficiente, por ejemplo el que corresponde a IMC, tendríamos que decir:

- Si el modelo tuviese capacidad predictiva ($R^2 > 0,80$) por cada unidad de IMC la Pam incrementa en 0,68 mmHg.
- Si el modelo tuviera fuerza asociativa, simplemente existe una asociación directa y significativa entre IMC y Pam (ya que el coeficiente es positivo).

El modelo final propuesto tiene un coeficiente de determinación de un 4,3%, que es bajo pero significativamente mayor a 0 ($p = 0,0000$), por lo tanto, la segunda conclusión es la más plausible.