

Regresión Lineal Múltiple

A long-exposure photograph of a winding road at night. The road curves through a dark, hilly landscape. Light trails from vehicles are visible, creating a sense of motion. The title 'Regresión Lineal Múltiple' is overlaid in white text.

Regresión Lineal Múltiple

- En la mayor parte de los problemas de investigación en que se aplica el análisis de regresión, se requiere más de una variable como predictor.
- La complejidad de la mayoría de los mecanismos científicos es tal que se necesita un modelo de regresión lineal múltiple.
- Es posible predecir la proporción de hogares con Necesidades Insatisfechas de los distritos de una provincia o estado mediante los valores tomados por otras variables como porcentaje de alfabetos, o porcentaje de hogares con Cobertura médica?

Ecuación de Regresión Lineal Múltiple

- *Dadas $x_1 \dots x_k$, k variables independientes, el modelo de regresión lineal múltiple es*

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

- *Para $i=1, 2, \dots, n$*
- *ε_i es el error del modelo, que se supone aleatorio y de media cero para todo i .*
- *Los parámetros deben estimarse a partir de los datos.*

Ecuación de Regresión Lineal Múltiple

- *La recta de regresión lineal múltiple estimada o ajustada es*

$$y_i = b_0 + b_1x_{1i} + \cdots + b_kx_{ki} + e_i$$

- *Para $i=1,2,\dots,n$*
- *$e_i = y_i - \hat{y}_i$ es el residuo y describe el error en el ajuste del modelo en cada punto.*

Estimación de los parámetros de la ecuación de Regresión Lineal Múltiple

- *Como en el caso de una única variable independiente, Los estimadores de los parámetros de la regresión, pueden hallarse mediante el método de mínimos cuadrados*
- *Se trata de encontrar los valores de b_0, b_1, \dots, b_k que minimicen*

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_{1i} - \dots - b_k x_{ki})^2$$

- *Es factible derivar fórmulas explícitas para estos estimadores mediante la resolución de este problema de minimización*
- *Observar que sólo se ha supuesto aleatoriedad de los errores y media cero*

Regresión Lineal múltiple

- *La suma total de cuadrados SST de la variable dependiente puede escribirse como*

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

- *En otras palabras*
 - *$SST = SSR + SSE$*

Regresión Lineal múltiple


- Se define el **coeficiente de determinación** como

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$


- El **coeficiente de determinación ajustado** es

$$R_a^2 = R^2 - \frac{p(1 - R^2)}{N - p - 1}$$

- El **coeficiente de correlación múltiple** es R



Una medida objetiva del ajuste aeróbico de una persona es el consumo de oxígeno en volumen por peso unitario del cuerpo, por unidad de tiempo. Se utilizaron 31 individuos en un experimento con el objeto de poder modelar el consumo de oxígeno mediante las siguientes variables.

- Edad.
 - Peso.
 - Tiempo en recorrer 3 km.
 - Pulso en reposo.
 - Pulso al final del ejercicio.
 - Pulso máximo durante el ejercicio.
- 

Resúmenes de casos^a

	Consumo de oxígeno	Edad	Peso	Tiempo en recorrer 3km	Pulso en reposo	Pulso al final del ejercicio	Pulso máximo durante el ejerc.
1	44.609	44.00	89.470	11.37	62.000	178.000	182.00
2	45.313	40.00	75.070	10.07	62.000	185.000	185.00
3	54.297	44.00	85.840	8.65	45.000	156.000	168.00
4	59.571	42.00	68.150	8.17	40.000	166.000	172.00
5	49.874	38.00	89.020	9.22	55.000	178.000	180.00
6	44.811	47.00	77.450	11.63	58.000	176.000	176.00
7	45.681	40.00	75.980	11.95	70.000	176.000	180.00
8	49.091	43.00	81.190	10.85	64.000	162.000	170.00
9	39.442	44.00	81.420	13.08	63.000	174.000	176.00
10	60.055	38.00	81.870	8.63	48.000	170.000	186.00
11	50.541	44.00	73.030	10.13	45.000	168.000	168.00
12	37.388	45.00	87.660	14.03	56.000	186.000	192.00
13	44.754	45.00	66.450	11.12	51.000	176.000	176.00
14	47.273	47.00	79.150	10.60	47.000	162.000	164.00
15	51.855	54.00	83.120	10.33	50.000	166.000	170.00
16	49.156	49.00	81.420	8.95	44.000	180.000	185.00
17	40.836	51.00	69.630	10.95	57.000	168.000	172.00
18	46.672	51.00	77.910	10.00	48.000	162.000	168.00
19	46.774	48.00	91.630	10.25	48.000	162.000	164.00
20	50.388	49.00	73.370	10.08	76.000	168.000	168.00
21	39.407	57.00	73.370	12.63	58.000	174.000	176.00
22	46.080	54.00	79.380	11.17	62.000	156.000	165.00
23	45.441	52.00	76.320	9.63	48.000	164.000	166.00
24	54.625	50.00	70.870	8.92	48.000	146.000	155.00
25	45.118	51.00	67.250	11.08	48.000	172.000	172.00
26	39.203	54.00	91.630	12.88	44.000	168.000	172.00
27	45.790	51.00	73.710	10.47	59.000	186.000	188.00
28	50.545	57.00	59.080	9.93	49.000	148.000	155.00
29	48.673	49.00	76.320	9.40	56.000	186.000	188.00
30	47.920	48.00	61.240	11.50	52.000	170.000	176.00
31	47.467	52.00	82.780	10.50	53.000	170.000	172.00
Total N	31	31	31	31	31	31	31

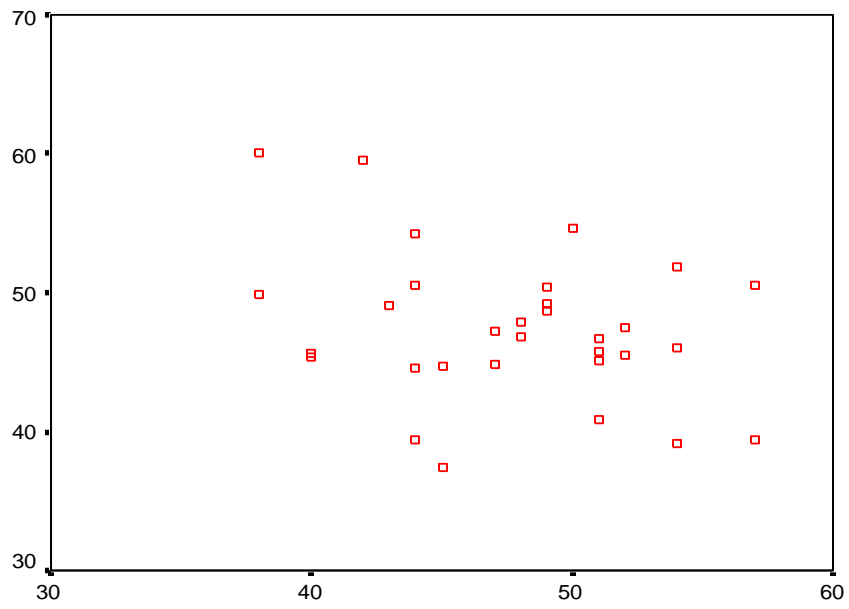
a. Limitado a los primeros 100 casos.

Correlaciones

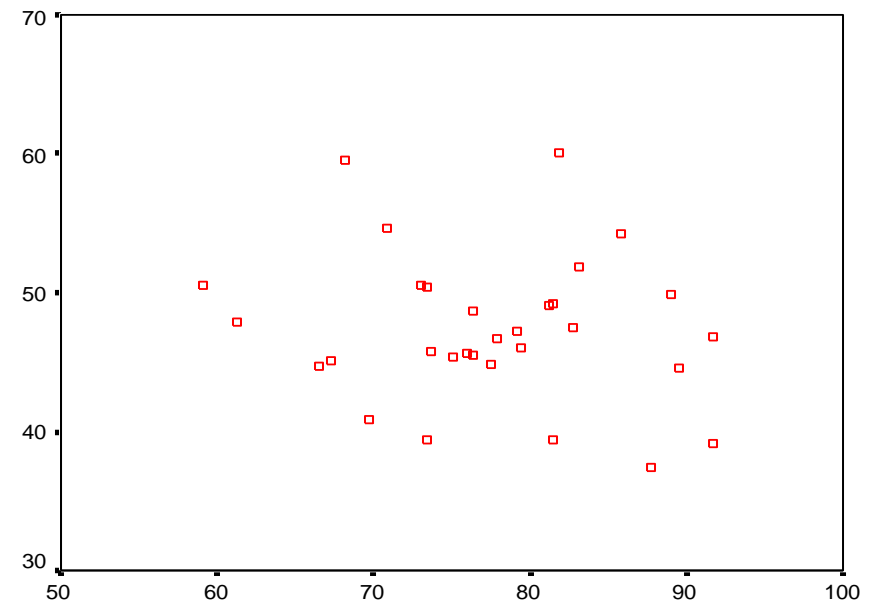
		Consumo de oxígeno	Edad	Peso	Tiempo en recorrer 3km	Pulso en reposo	Pulso al final del ejercicio	Pulso máximo durante el ejerc.
Consumo de oxígeno	Correlación de Pearson	1	-.305	-.163	-.862**	-.346	-.398*	-.237
	Sig. (bilateral)	.	.096	.382	.000	.056	.027	.200
	N	31	31	31	31	31	31	31
Edad	Correlación de Pearson	-.305	1	-.234	.189	-.142	-.338	-.433*
	Sig. (bilateral)	.096	.	.206	.309	.447	.063	.015
	N	31	31	31	31	31	31	31
Peso	Correlación de Pearson	-.163	-.234	1	.144	.023	.182	.249
	Sig. (bilateral)	.382	.206	.	.441	.904	.328	.176
	N	31	31	31	31	31	31	31
Tiempo en recorrer 3km	Correlación de Pearson	-.862**	.189	.144	1	.401*	.314	.226
	Sig. (bilateral)	.000	.309	.441	.	.026	.086	.221
	N	31	31	31	31	31	31	31
Pulso en reposo	Correlación de Pearson	-.346	-.142	.023	.401*	1	.318	.258
	Sig. (bilateral)	.056	.447	.904	.026	.	.081	.162
	N	31	31	31	31	31	31	31
Pulso al final del ejercicio	Correlación de Pearson	-.398*	-.338	.182	.314	.318	1	.930**
	Sig. (bilateral)	.027	.063	.328	.086	.081	.	.000
	N	31	31	31	31	31	31	31
Pulso máximo durante el ejerc.	Correlación de Pearson	-.237	-.433*	.249	.226	.258	.930**	1
	Sig. (bilateral)	.200	.015	.176	.221	.162	.000	.
	N	31	31	31	31	31	31	31

** . La correlación es significativa al nivel 0,01 (bilateral).

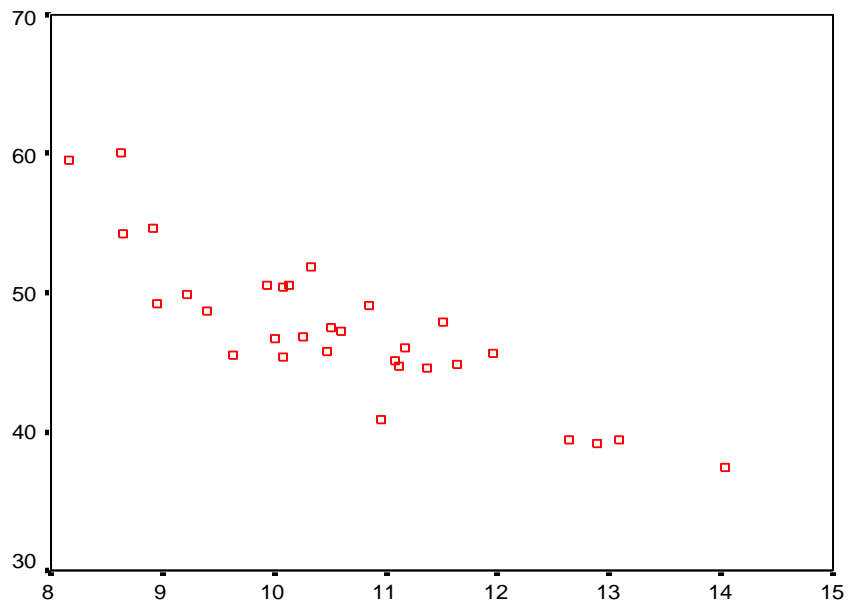
* . La correlación es significativa al nivel 0,05 (bilateral).



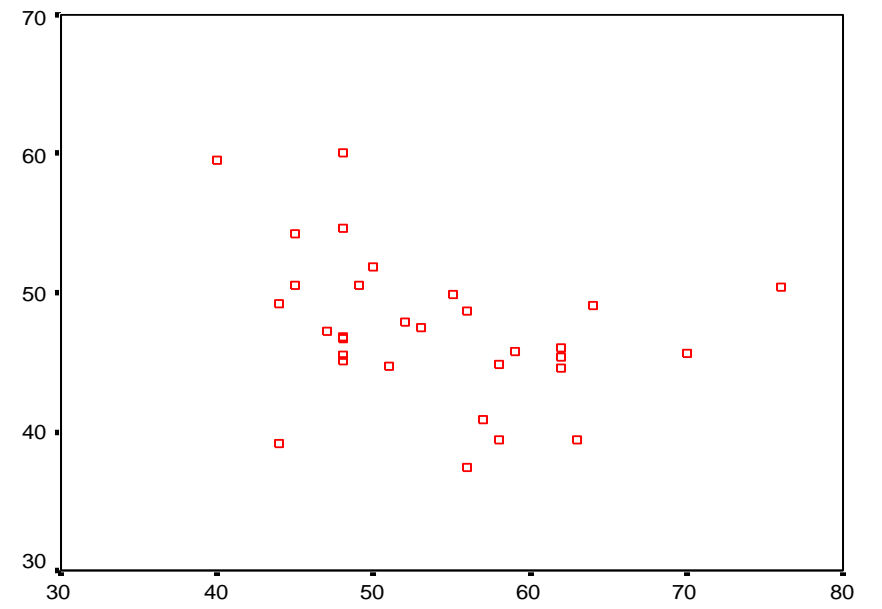
Edad



Peso



Tiempo en recorrer 3km



Pulso en reposo



Pulso al final del ejercicio



Pulso máximo durante el ejerc.

Model Summary

R	0.921	RMSE	2.322
R-Squared	0.848	Coef. Var	4.901
Adj. R-Squared	0.810	MSE	5.392
Pred R-Squared	0.763	MAE	1.437

RMSE: Root Mean Square Error

MSE: Mean Square Error


MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	721.974	6	120.329	22.316	0.0000
Residual	129.407	24	5.392		
Total	851.382	30			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	102.238	12.453		8.210	0.000	76.537	127.940
x1	-0.220	0.100	-0.215	-2.208	0.037	-0.425	-0.014
x2	-0.072	0.055	-0.113	-1.324	0.198	-0.185	0.040
x3	-2.681	0.375	-0.698	-7.150	0.000	-3.454	-1.907
x4	-0.001	0.059	-0.001	-0.014	0.989	-0.122	0.120
x5	-0.373	0.121	-0.718	-3.092	0.005	-0.622	-0.124
x6	0.305	0.137	0.524	2.221	0.036	0.022	0.588



Selección de predictores

La búsqueda del mejor modelo

Selección de modelos

El mejor modelo de regresión depende de los objetivos que se tengan

No siempre el mejor modelo es el que tiene más predictores.

Muchas veces no existe el mejor modelo...

Por eso existen variados criterios de comparación entre modelos y es tan importante la selección de predictores.

Selección de modelos

- *La inclusión de cualquier variable en la ecuación de regresión lineal incrementará SSR y, por ende, disminuirá SSE.*
- *El uso de variables no importantes puede reducir la efectividad de la ecuación de regresión, al incrementar la varianza de la respuesta estimada.*
- *En consecuencia, se deberá contar con un criterio que permita determinar si el incremento en SSR es suficientemente importante como para justificar su inclusión*

Selección de modelos

- *Un criterio para esto es evaluar la cantidad de variación en la respuesta atribuida a cada variable que ingresa en el modelo*
- *Este concepto es la base de los métodos secuenciales de selección de modelos*

A large orange circle is positioned on the left side of the slide, partially cut off by the edge.

Criterios de comparación de modelos

- Algunos criterios de selección de modelos
 - R cuadrado ajustado
 - Cp de Mallows
 - AIC (Akaike Information Criterion)
 - BIC (Bayesian Information Criterion)



Coeficiente de determinación

- Se define el **coeficiente de determinación** como

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- El **coeficiente de determinación ajustado** es

$$R_a^2 = R^2 - \frac{p(1 - R^2)}{N - p - 1}$$

- El **coeficiente de correlación múltiple** es R

Cp de Mallows

- El Cp de Mallows es un indicador del sesgo que se introduce al predecir la variable dependiente con un modelo mal especificado.
- Un modelo adecuado tiene un Cp igual o menor a la cantidad de parametros del modelo.
- Estrategia de uso de Cp
 - Identificar modelos con Cp cercano a $k+1$ y quedarse con el más sencillo de identificar, donde k es el número de variables

Estimación de los parámetros en Regresión Lineal Múltiple

- *En el método de mínimos cuadrados, se trata de encontrar los valores de b_0, b_1, \dots, b_k que minimicen*

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_{1i} - \dots - b_k x_{ki})^2$$

- *En el caso del método de máxima verosimilitud, se trata de encontrar los valores de b_0, b_1, \dots, b_k que minimicen la función de probabilidad conjunta de los datos de la muestra, llamada función de verosimilitud*

Indicadores de Akaike

- AIC provee una medida de selección de modelos. AIC estima la calidad de un modelo de acuerdo a dos criterios: El nivel de precisión de sus estimadores, y la cantidad de variables que contiene.

$$AIC = -2 \log(\text{función de verosimilitud}) + 2 * p$$

- SBC o BIC es una alternativa al AIC de Akaike. Se diferencia del AIC por su criterio de penalización de los modelos

$$BIC = -2 \log(\text{función de verosimilitud}) + \log(n) * 2 * p$$

- p es la cantidad de parámetros del modelo



Selección de modelos

- Dos estrategias de selección automática de predictores
 - Ajustar todos los posibles modelos (all possible subsets)
 - Método de selección de a pasos (stepwise methods)



Métodos secuenciales de selección de modelos

- *Forward*

`ols_step_forward_p(model, penter = 0.3, progress = FALSE, details = FALSE)`

- *Backward*

`ols_step_backward_p(model, prem = 0.3, progress = FALSE, details = FALSE)`

- *Stepwise*

`ols_step_both_p(model, pent = 0.1, prem = 0.3, progress = FALSE, details = FALSE)`

- *Todas las funciones son del paquete **olsrr***

Métodos secuenciales de selección de modelos

Método forward

- *Paso 1: Se selecciona la variable que da la SSR más grande al realizar una regresión lineal simple (x_1)*
- *Paso 2: Se selecciona la variable que, cuando es insertada en el modelo, da el mayor incremento de SSR, en presencia de x_1 (x_2)*
- *Se continúa este proceso hasta que ninguna de las variables produce un incremento en SSR en presencia de las que ya están incorporadas*

Métodos secuenciales de selección de modelos

Método backward

- *Utiliza el mismo criterio que el anterior salvo que se inicia con un modelo que contiene todas las variables independientes consideradas.*
- *Paso 1: Se selecciona y se elimina del modelo aquella variable que da el valor más pequeño SSR ajustada por las otras (x_1)*
- *Paso 2: Se ajusta el modelo de regresión con las variables restantes y se repite el paso 1*
- *Se continúa de esta forma hasta que no queden en la ecuación variables cuya contribución al SSR no sea significativa*

Métodos secuenciales de selección de modelos

Método stepwise

- *Es una modificación del forward, que incorpora en cada etapa un chequeo de la eficacia de las variables ingresadas en el modelo en las etapas anteriores. Este chequeo puede derivar en la eliminación de alguna de estas variables*



multicolinealidad

Multicolinealidad

Es el fenómeno que se produce cuando los predictores están correlacionados entre sí.

Efectos de la colinealidad

- Limita el tamaño del R^2 por que los predictores correlacionados “explican” la misma variación
- Dificulta la determinación de la importancia de los predictores
- Incrementa la varianza de los estimadores de los parámetros

Multicolinealidad


Diagnósticos de multicolinealidad

- Examinar la correlación entre los predictores
- Variance inflation factor $VIF = \frac{1}{1-R_j^2}$
 - Indica cuándo un predictor tiene una fuerte asociación lineal con el resto de los predictores
 - $VIF > 10$ indica multicolinealidad
- Tolerancia $Tol = \frac{1}{VIF}$
 - $Tol < 0.10$ indica multicolinealidad



Multicolinealidad

Soluciones

- Combinar predictores que están altamente correlacionados
 - Hacer componentes principales y utilizar como predictores las nuevas variables
 - Regresión de Ridge
- 

- En el ejemplo de correr 3 km.

Tolerance and Variance Inflation
Factor

	Variables	Tolerance	VIF
1	x1	0.6672145	1.498768
2	x2	0.8668311	1.153627
3	x3	0.6643874	1.505146
4	x4	0.7599631	1.315853
5	x5	0.1174186	8.516537
6	x6	0.1136534	8.798680

Multicolinealidad
- Ejemplo