Los datos siguientes contienen indicadores demográficos y criminales sobre 47 estados de EEUU. Los datos fueron recogidos por la *Uniform Crime Report* del FBI y otras agencies del gobierno para determinar cómo la variable dependiente **tasa de criminalidad**, ***R***, depende de otras variables medidas en el estudio.

Las variables son:

R: tasa de criminalidad: número de denuncias por millón de habitantes,
Age: número de varones entre 14-24 por 1000 habitantes,
Ed: media del número de años de educación x 10 para personas de 25 años o mayores,
Ex0: inversiones en seguridad (per cápita) realizada por los gobiernos local y estatal en 1960,
Ex1: inversiones en seguridad (per cápita) realizada por los gobiernos local y estatal en 1959,
LF: tasa de mano de obra por 1000 varones urbanos de entre 14-24 años,
M: número de varones por 1000 mujeres,
N: tamaño de la población del estado en cientos de miles,
NW: número de no-blancos por 1000 habitantes,
U1: tasa de paro en varones urbanos por 1000 de entre 14-24 años,
U2: tasa de paro en varones urbanos por 1000 de entre 35-39 años,
W: valor mediano de bienes y activos transferibles o renta familiar en decenas de dólares,
X: número de familias por 1000 con una renta por debajo de la mitad de la renta mediana.

Problemas que nos vamos a encontrar:

- alta correlación entre Ex0 y Ex1, entre W y X, también entre U1 y U2. (provocará que las estimaciones de los coeficientes sean inestables; quitar una variable del modelo hará que las otras cambien drásticamente)

- la relación causal entre Ex0 y R no es demasiado clara: ¿las inversiones del gobierno en seguridad afectan a la tasa de crimen o bien las inversiones del gobierno se producen porque aumenta la tasa de crimen?

Multicolinealidad

a) Factores de incremento de la varianza:

La diagonal de la inversa de la matriz de correlaciones entre las variables explicativas es:

diag(inv (**R**)) = (2.6968, 5.0126, **94.6293**, **98.3901**, 2.7457, 3.5042, 2.3104, 3.1828, 5.1316, 4.8052, **9.6640**, 7.3547)

Existen valores mayores que 10, $FIV_{(3)}$= **94.6293**, $FIV_{(4)}$= **98.3901**, lo que indica una alta multicolinealidad.

b) Índice de condicionamiento:

El índice de condicionamiento de la matriz **R** es:

$$cond(R) = \sqrt{\frac{5.0534}{0.0052}} = 31.0729 > 30 \Rightarrow \text{alta multicolinealidad}$$

Coclusión: Tendremos problemas de multicolinealidad si consideramos **todas las variables.**

Selección de variables.

La selección de un conjunto reducido de variables explicativas es un problema complicado:

- número demasiado pequeño de variables → modelo poco potente, sesgo en los coeficientes de regresión y en las predicciones (porque los errores calculados con los datos observados pueden contener efectos de las variables desechadas),
- número muy grande de variables → poca utilidad práctica del modelo, aumento de la varianza de los estimadores de los parámetros (a pesar de tener buen ajuste).

Métodos de selección de variables:

1. **Coeficiente de determinación ajustado,**
2. **Criterio C_p de Mallows,**
3. **Selección paso a paso.**

1. Coeficiente de determinación ajustado.

Calcular todos los coeficientes de determinación ajustados de todos los modelos posibles con la combinación de cualquier número de variables explicativas.

Seleccionar el modelo cuyo R^2-ajustado sea mayor.

2. Criterio C_P de Mallows.

Se trata de hallar el mejor modelo con P variables explicativas, incluida la constante, utilizando el estadístico de Mallows

$$C_P = \frac{VNE_P}{s_R^2} - (n - 2P)$$

donde VNE_P es la suma de cuadrados residual del modelo particular (es decir, del modelo con P variables explicativas, incluida la constante) y s_R^2 es la varianza residual del modelo completo ($P=k+1$).
Para el modelo completo $P=k+1$, el estadístico de Mallows es:

$$C_P = \frac{VNE}{s_R^2} - (n - 2(k+1)) = n - k - 1 - (n - 2(k+1)) = k + 1$$

(puesto que $s_R^2 = VNE/(n-k-1)$.)

Para todo modelo no completo, si el modelo es adecuado, $E(C_P)=P$.
Por tanto, es recomendable seleccionar el modelo cuya C_P sea aproximadamente P.

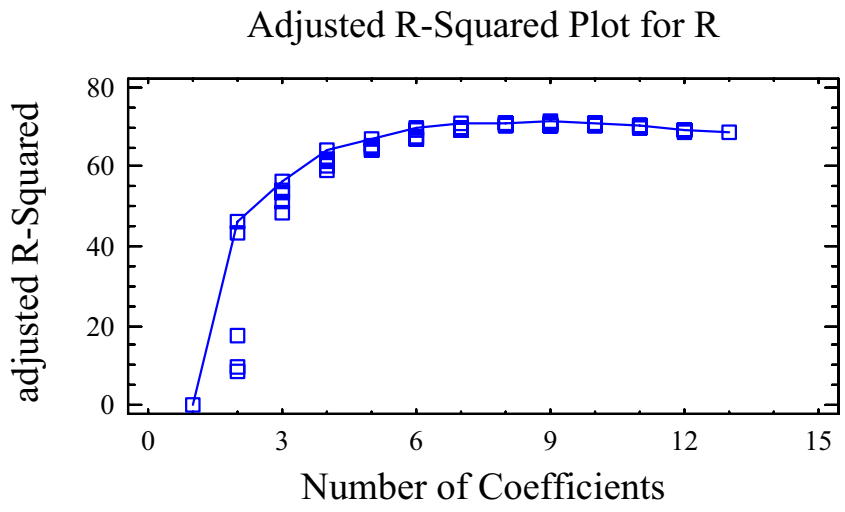
Regression Model Selection

Dependent variable: R
Independent variables:
A=Age, B=Ed, C=Ex0, D=Ex1, E=LF, F=M, G=N, H=NW, I=U1, J=U2, K=W, L=X

Number of models fit: 4096

Models with Largest Adjusted R-Squared

MSE	R-Squared	**Adjusted R-Squared**	Cp	Included Variables
427,67	76,3819	**71,4097**	5,47385	**ABCFIJKL** (**8 variables**)
433,777	74,7838	71,0014	3,80642	ABCJKL
434,422	75,3777	70,9583	4,93965	ABCIJKL
434,594	76,6311	70,9468	7,1101	ABCDFIJKL
436,166	75,2788	70,8417	5,0839	ABCFIJL
437,985	75,8122	70,7201	6,30531	ABCDIJKL
438,255	76,4343	70,702	7,39741	ABCFHIJKL
438,457	75,149	70,6885	5,27347	ABCDJKL
438,604	76,4154	70,6787	7,42486	ABCFGIJKL
438,946	75,7592	70,6558	6,38278	ABCGIJKL
438,984	75,1191	70,6533	5,31708	ABCEJKL



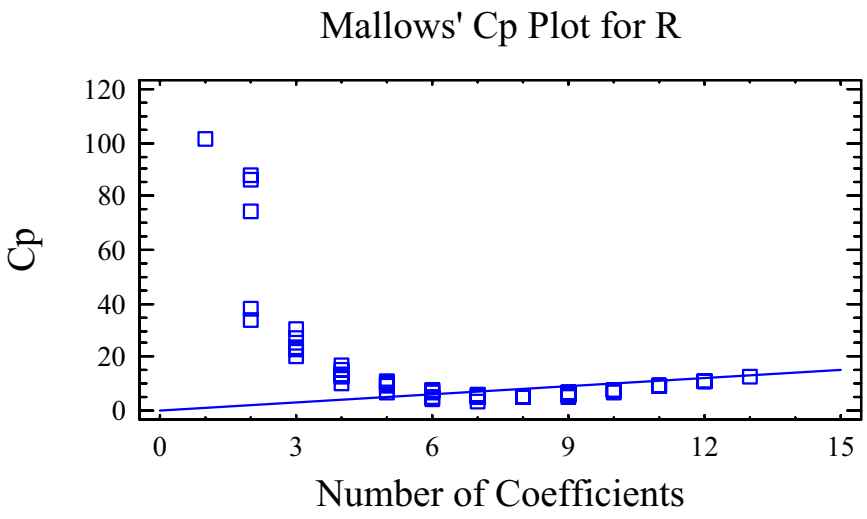
Regression Model Selection

Dependent variable: R
Independent variables:
A=Age, B=Ed, C=Ex0, D=Ex1, E=LF, F=M, G=N, H=NW, I=U1, J=U2, K=W, L=X

Number of models fit: 4096

Models with Smallest Cp

MSE	R-Squared	Adjusted R-Squared	**Cp**	Included Variables
433,777	74,7838	71,0014	3,80642	ABCJKL
453,747	72,9635	**69,6663**	**4,4635**	**ABCJL**
434,422	75,3777	70,9583	4,93965	ABCIJKL
436,166	75,2788	70,8417	5,0839	ABCFIJL
448,983	73,8999	69,9848	5,09669	ABCIJL
462,921	72,4169	69,0531	**5,26131**	**ABCKL**
438,457	75,149	70,6885	5,27347	ABCDJKL
438,984	75,1191	70,6533	5,31708	ABCEJKL
440,039	75,0593	70,5828	5,40432	ABCGJKL
427,67	76,3819	71,4097	5,47385	ABCFIJKL
454,617	73,5724	69,6082	5,57472	ABCEJL



3. Selección paso a paso.

El procedimiento se puede realizar hacia adelante (**forward stepwise)** o hacia atrás (**backward stepwise**), seleccionando las variables una a una e incorporándolas desde el modelo inicial o eliminándolas desde el modelo completo en función de su contribución al modelo.

Inconveniente: este método puede conducir a modelos distintos y no necesariamente óptimos.

En la selección hacia adelante se incorpora como primera variable la de mayor *F* de significación de la regresión simple. La segunda variable se selecciona por su mayor contribución al modelo que ya contiene la primera variable del paso anterior y así sucesivamente.

Multiple Regression Analysis (Stepwise regression: forward selection)

Dependent variable: R

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-524,374	95,1156	-5,51302	0,0000
Age	1,01982	0,353203	2,88736	0,0062
Ed	2,03077	0,474189	4,28262	0,0001
Ex0	1,23312	0,141635	8,70636	0,0000
U2	0,913608	0,434092	2,10464	0,0415
X	0,634926	0,146846	4,32375	0,0001

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	50205,6	5	10041,1	22,13	0,0000
Residual	18603,6	41	453,747		
Total (Corr.)	68809,3	46			

R-squared = 72,9635% Standard Error of Est. = 21,3013
R-squared (adjusted for d.f.) = **69,6663**% Mean absolute error = 14,9425

¿Persisten los problemas de multicolinealidad?

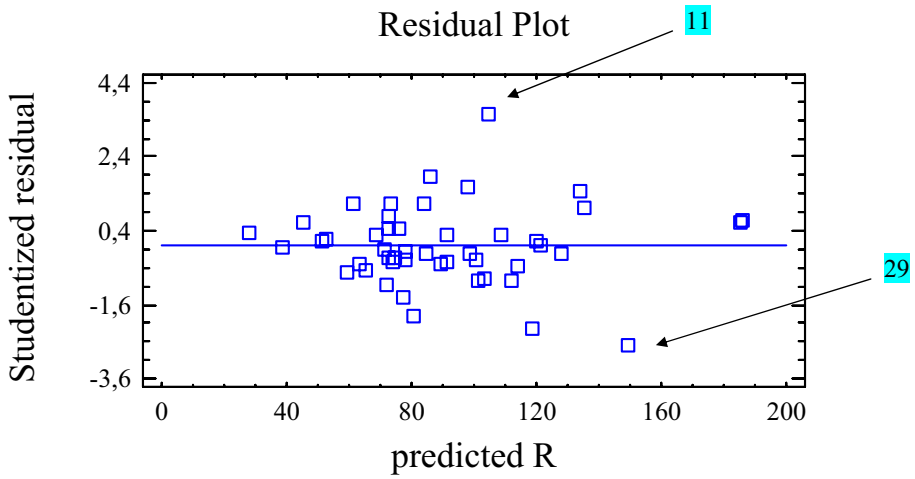
No, puesto que no se observa ningún valor superior a 10 en

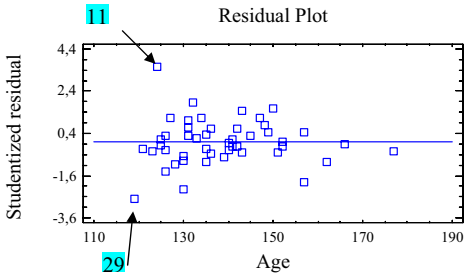
$$\text{diag}\,(\mathbf{R}^{-1}) = (1.9976, 2.8528, 1.7962, 1.3626, 3.4796)$$

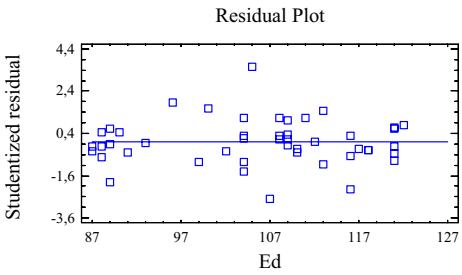
y el índice de condicionamiento es:

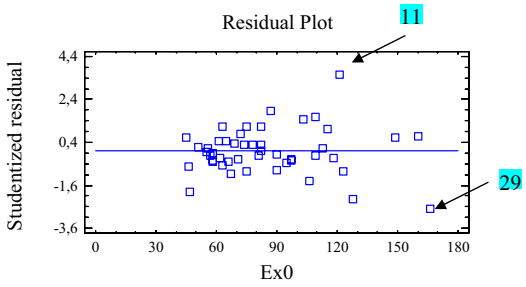
$$(2.7935/0.1941)^{1/2} = 3.7933 < 10 \rightarrow \mathbf{R} \text{ está bien definida.}$$

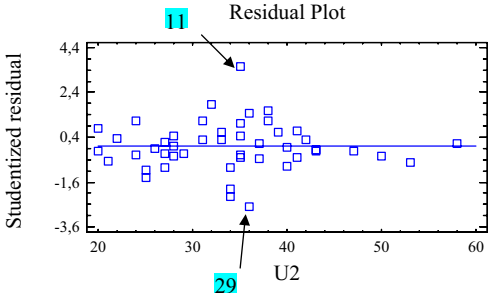
Gráficos de residuos:

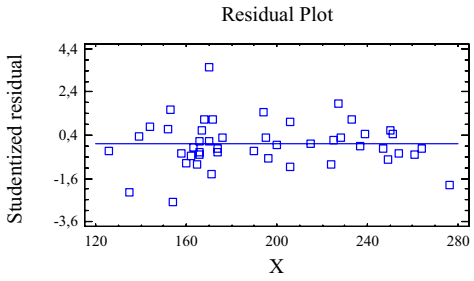






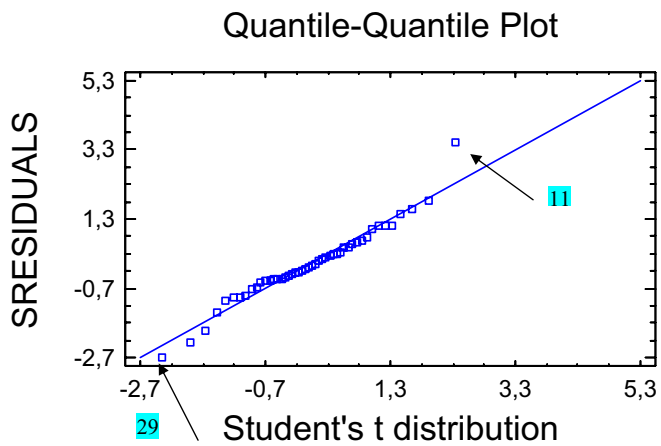






Influential Points

Row	Leverage (v_{ii})	Average leverage of single data point
11	0,0999826	$\bar{v}$=0.12766
29	**0,267829**	
36	0,22159	
		$2\bar{v}=2(0.12766)=0.2555$



Datos atípicos

Row	SRESIDUAL (t_i)
11	**3,52266**
29	-2,66671

Student's ***t (α=0.05)***:
$t_{n-k-2} = t_{40} = 2.70$

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0,087459	0,61149	>=0.10
Cramer-Von Mises W^2	0,070591	0,063679	>=0.10
Anderson-Darling A^2	0,439835	0,439835	>=0.10

El modelo estimado es:

$$R = -524{,}37 + 1{,}02*Age + 2{,}03*Ed + 1{,}23*Ex0 + 0{,}91*U2 + 0{,}63*X$$

Las variables:

- Age (nº de varones jóvenes por 1000hab),
- U2 (tasa de paro en varones urbanos jóvenes por 1000hab),
- X (nº de familias por 1000 con una renta por debajo de la mitad de la renta mediana),

contribuyen positivamente al incremento de la tasa de criminalidad.

Anomalías:

- Un aumento de la variable Ed (media del número de años de educación) causa un aumento de la tasa de criminalidad !!!???
  (Posiblemente existe una variable, que no se ha tenido en cuenta en este estudio, que causa que tanto la educación como la tasa de crimen aumenten simultáneamente).
- La causalidad de la variable Ex0 (inversiones en seguridad) es discutible.