Ejemplo de error de especificación.

Consideremos el siguiente conjunto de datos:

y	x_1	x_2
24,2	10	2
84	20	1
185,8	30	4
34,7	12	3
41	14	2
58	16	2
69	18	4
98,4	22	5
127	24	1
133	26	3
159.2	28	5

Supongamos que hemos especificado el modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Calculamos las medias, varianzas y desviaciones:

	x1	x2	y
Average	20,0	2,90909	92,2091
Variance	44,0	2,09091	2849,09
Standard deviation	6,63325	1,446	53,3769

Y la matriz de correlaciones:

	x1	x2	y
x1		**0,4170** (11) *0,2019*	**0,9882** (11) *0,0000*
x2	**0,4170** (11) *0,2019*		**0,4107** (11) *0,2095*
y	**0,9882** (11) *0,0000*	**0,4107** (11) *0,2095*	

donde el valor en negrita es el coef. de correlación, entre paréntesis se encuentra el tamaño muestral, y en cursiva el p-valor.

Multicolinealidad?

$$R = \begin{pmatrix} 1 & 0.4170 \\ 0.4170 & 1 \end{pmatrix} \qquad R^{-1} = \begin{pmatrix} 1.2105 & -0.5048 \\ -0.5048 & 1.2105 \end{pmatrix}$$

$FIV_{(1)}=FIV_{(2)}=1.2105<10$

No parece que vayamos a tener problemas de multicolinealidad.

La estimación del modelo es:

Multiple Regression Analysis
Dependent variable: y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-66,7601	9,45447	-7,06122	0,0001
x1	7,95742	0,479656	16,5899	0,0000
x2	-0,0615789	2,20033	-0,0279862	0,9784

$\hat{y}$ = -66.76 + 7.96 x_1 – 0.06 x_2

De esta tabla concluimos que x_1 es significativa pero que x_2 no lo es, puesto que:

$$\left|t_{\exp}(\beta_1)\right| = \left|\frac{7.95742}{0.479656}\right| = \left|16.5899\right| > 2.31 = t_8$$

$$\left|t_{\exp}(\beta_2)\right| = \left|\frac{-0.0615789}{2.20033}\right| = \left|-0.0279862\right| < 2.31 = t_8$$

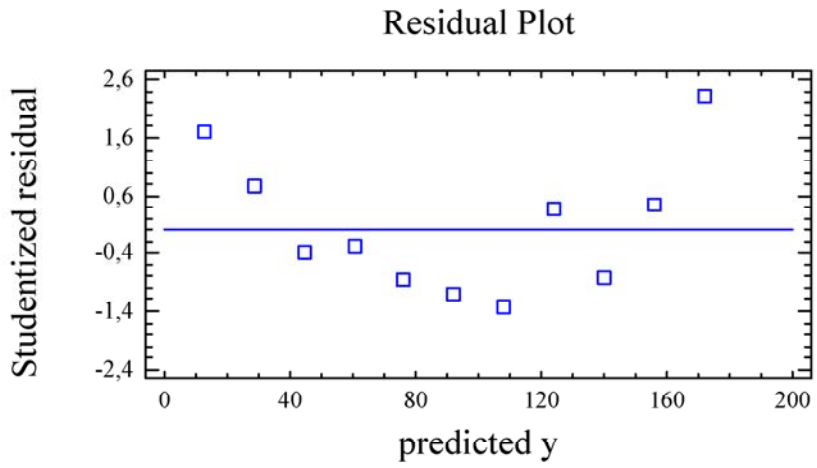
Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	27821,9	2	13910,9	166,35	0,0000
Residual	669,002	8	83,6253		
Total (Corr.)	28490,9	10			

R-squared = **97,6519** percent
R-squared (adjusted for d.f.) = 97,0648 percent
Standard Error of Est. = 9,14469
Mean absolute error = 6,9187
Durbin-Watson statistic = 2,15374

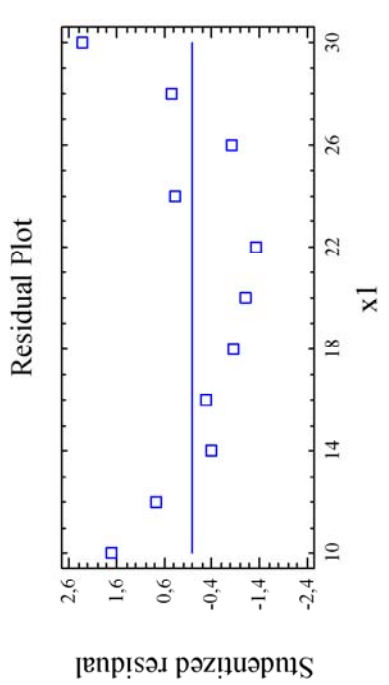
El análisis de la varianza nos dice que el modelo de regresión explica un 97.65% de la variabilidad de la respuesta ***y***. De acuerdo con estos resultados parece que la variable x_2 debería eliminarse del modelo y que la variable x_1 es muy significativa.

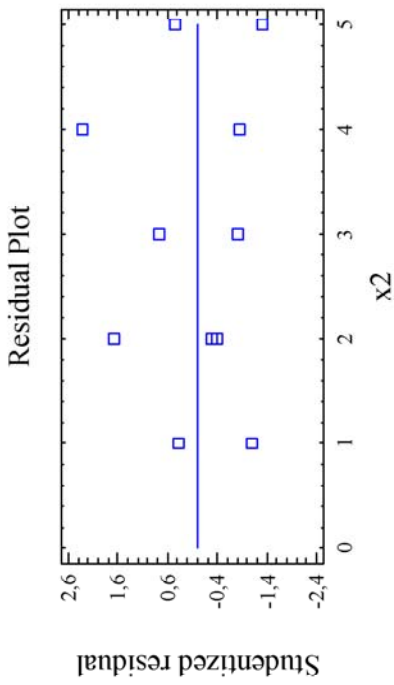
Estudiemos los residuos:



Este gráfico muestra claramente un error de especificación: los residuos toman una forma parabólica, señalando la necesidad de transformar alguna o varias de las variables.

Para detectar qué variable explicativa produce este efecto realizamos el gráfico de los residuos frente a x_1 y frente a x_2





La estructura de los residuos frente a x_1 es análoga a la observada respecto a $\hat{y}$, indicando la necesidad de transformar esta variable "estirando la escala de x" o bien "comprimiendo la escala de y".

Una solución intuitiva es tomar x_1^2 en lugar de x_1 como variable explicativa.
Por tanto, planteamos el modelo: $y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2$

Multiple Regression Analysis

Dependent variable: y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	6,96679	2,68524	2,59447	0,0319
x1^2	0,201616	0,00468017	43,0788	0,0000
x2	1,19239	0,86711	-1,37513	0,2064

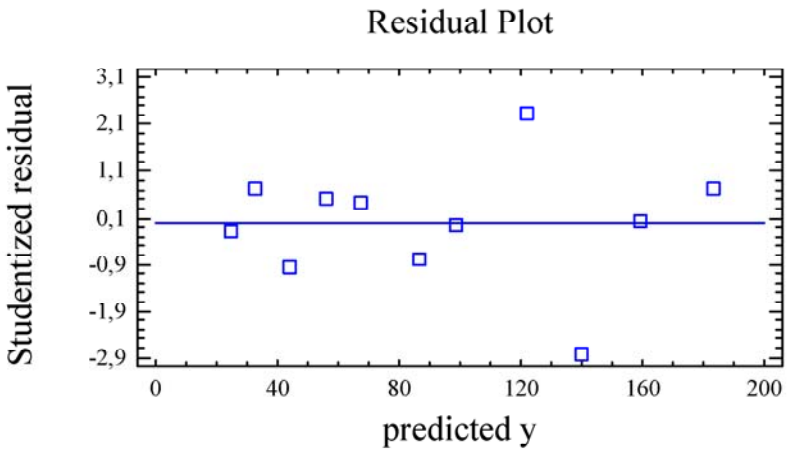
El modelo estimado es $\hat{y} = 6.97 + 0.20\, x_1^2 - 1.19\, x_2$

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	28389,2	2	14194,6	1117,00	0,0000
Residual	101,663	8	12,7078		
Total (Corr.)	28490,9	10			

R-squared = **99,6432** percent
R-squared (adjusted for d.f.) = 99,554 percent
Standard Error of Est. = 3,5648
Mean absolute error = 2,35133

Ahora el gráfico de los residuos frente a ŷ no muestra ninguna forma parabólica:



y los gráficos de los residuos frente a x_1 y frente a x_2 son: