UNIVERSIDAD
COMPLUTENSE
MADRID

Proyectos de Innovación y Mejora de la Calidad Docente

Vicerrectorado de Evaluación de la Calidad

Convocatoria 2015, Proyecto núm. 164

«Videotutoriales de estadística aplicada a las Ciencias Sociales: un recurso formativo emergente en las actividades de enseñanza-aprendizaje»

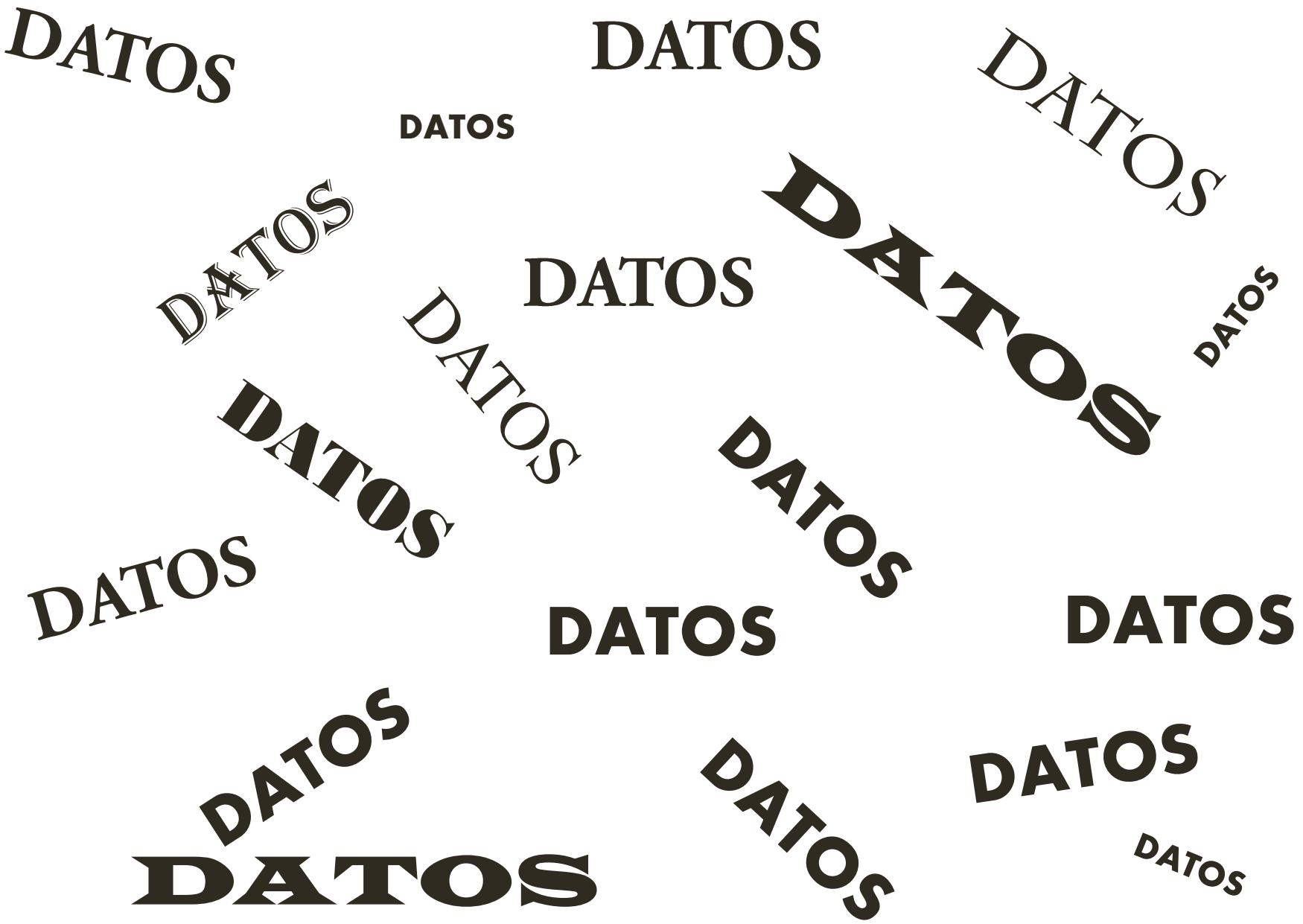
Departamento de SOCIOLOGIA IV

Metodología de la investigación y Teoría de la Comunicación

Proyectos de Innovación y Mejora de la Calidad Docente
Vicerrectorado de Evaluación de la Calidad

ANÁLISIS DE LA RELACIÓN ENTRE VARIABLES CUANTITATIVAS MEDIANTE REGRESIÓN LINEAL MÚLTIPLE

Carlos Montes Botella (Dto. Sociología IV – UCM)


DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS
DATOS

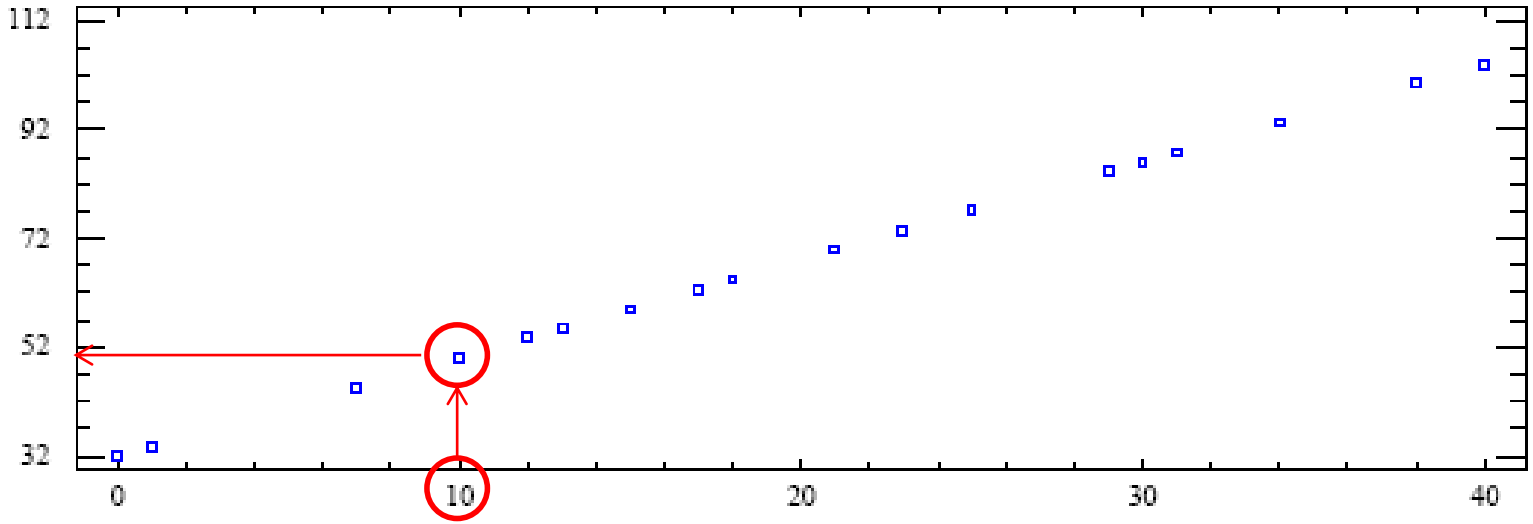
Independiente

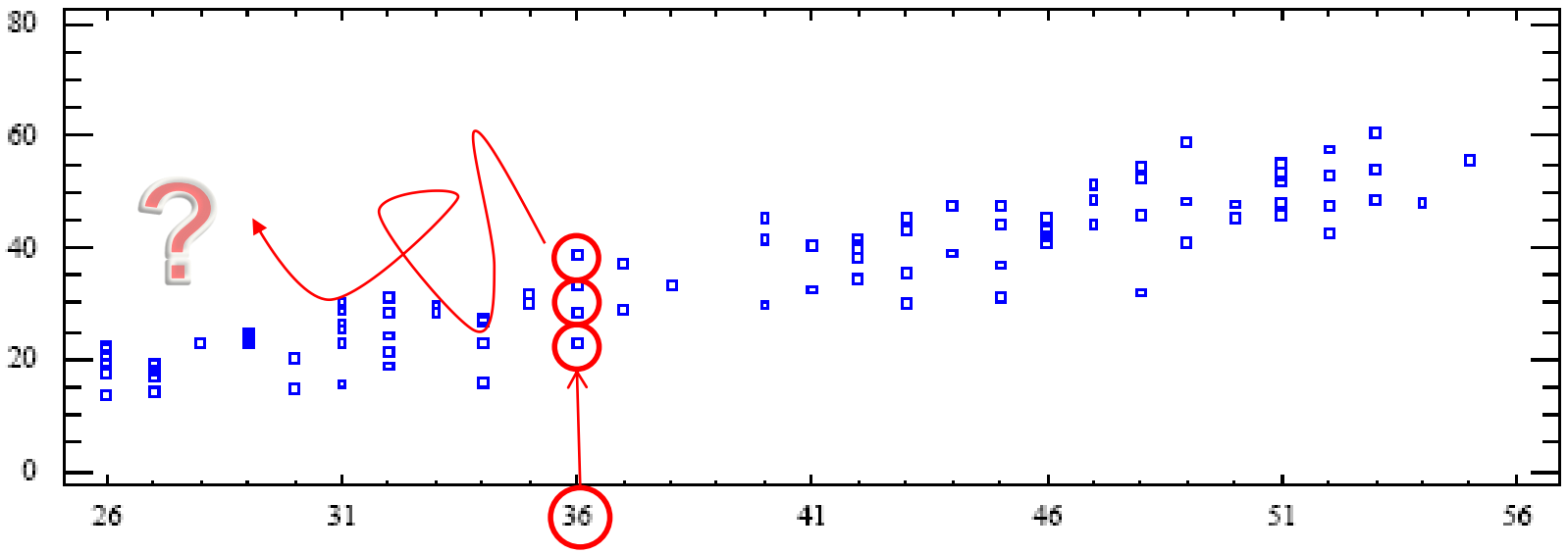
Eje horizontal

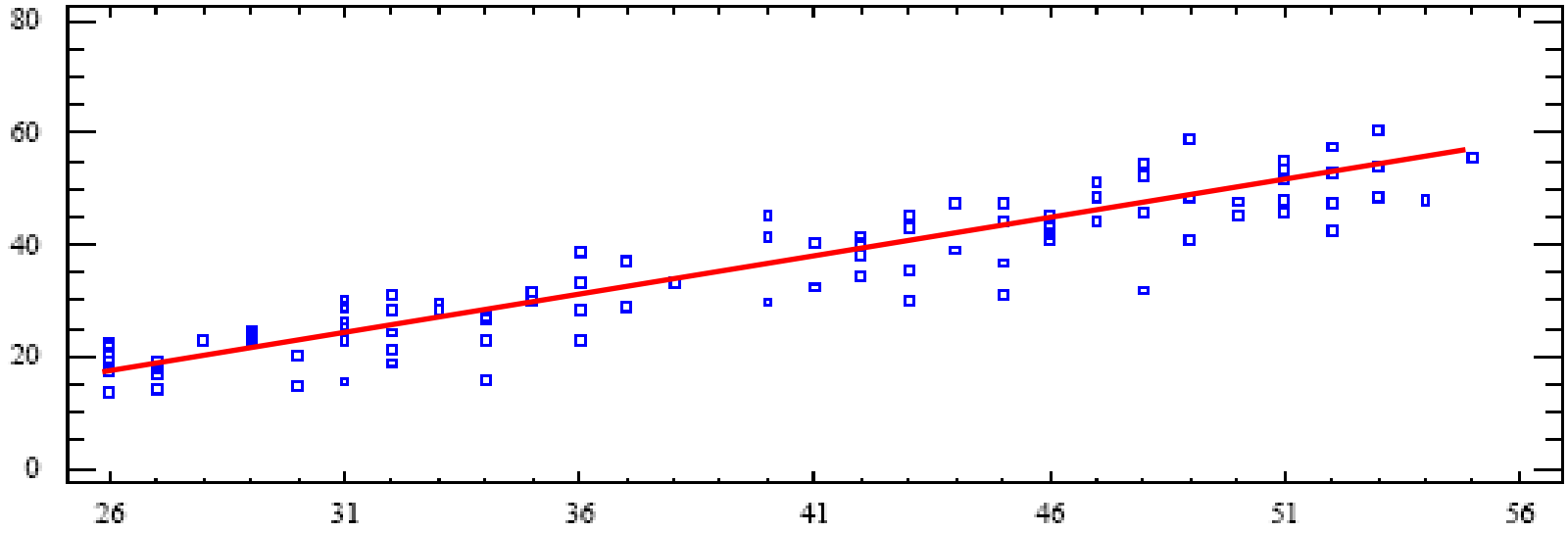
Dependiente

Eje vertical

Regresión


112
92
72
52
32
0
10
20
30
40


80
60
40
20
0
26
31
36
41
46
51
56



$$y_i = a + bx_i + e_i$$

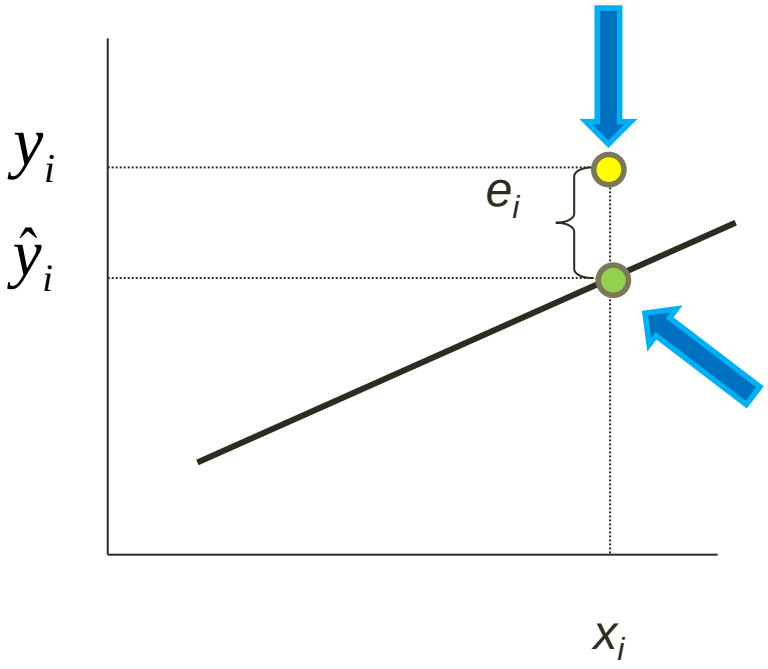
Variable respuesta

error

Variable explicativa

Para cada x_i tendremos

- valor real y_i
- valor sobre la recta de regresión $\hat{y}_i$

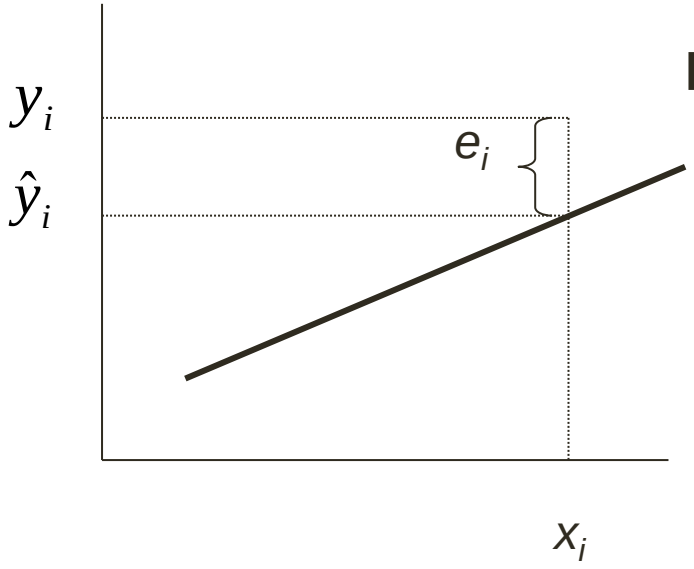


$$y_i = \hat{y}_i + e_i$$

$$e_i = y_i - \hat{y}_i$$

residuo

Posible criterio de construcción: minimizar la suma de los errores.



$$e_i = y_i - \hat{y}_i$$

$$\min \sum_{i=1}^{n} e_i = \min \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)$$

Para evitar la influencia de los signos:

$$\min \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Método de los mínimos cuadrados

Carl Friedrich Gauss (1777-1855)

Podemos preguntarnos
el valor de una variable respuesta
a partir de más de una variable explicativa.

Modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i$$

variables independientes o explicativas

También se le denomina “recta de regresión” aunque no sea una recta, sino un hiperplano.

Hipótesis del modelo:

1: Linealidad

2: Homocedasticidad

3: Independencia

4: Normalidad

Si no se cumplen las hipótesis,
hay que realizar una transformación de los datos

$$\ln$$

$$x^2$$

$$\sqrt{x}$$

$$e^x$$

¿Cómo valorar el ajuste?

El coeficiente R^2 (coeficiente de determinación) indica el porcentaje de la variación de Y que es explicada por X.

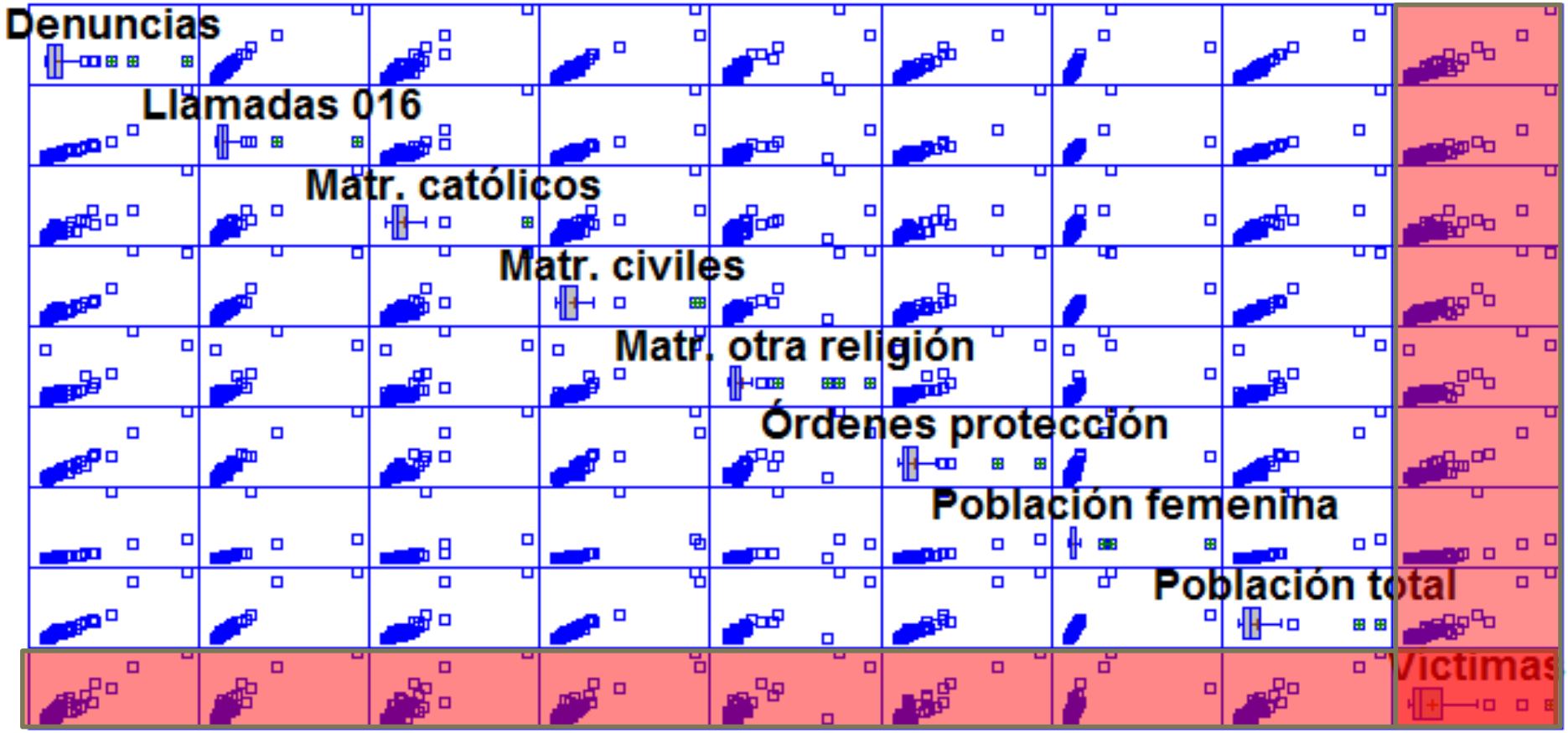
A medida que aumenta el número de variables, R^2 puede aumentar aunque las variables no sean significativas.

Por ello se define el coeficiente de determinación corregido o ajustado.

Víctimas mortales por violencia de género

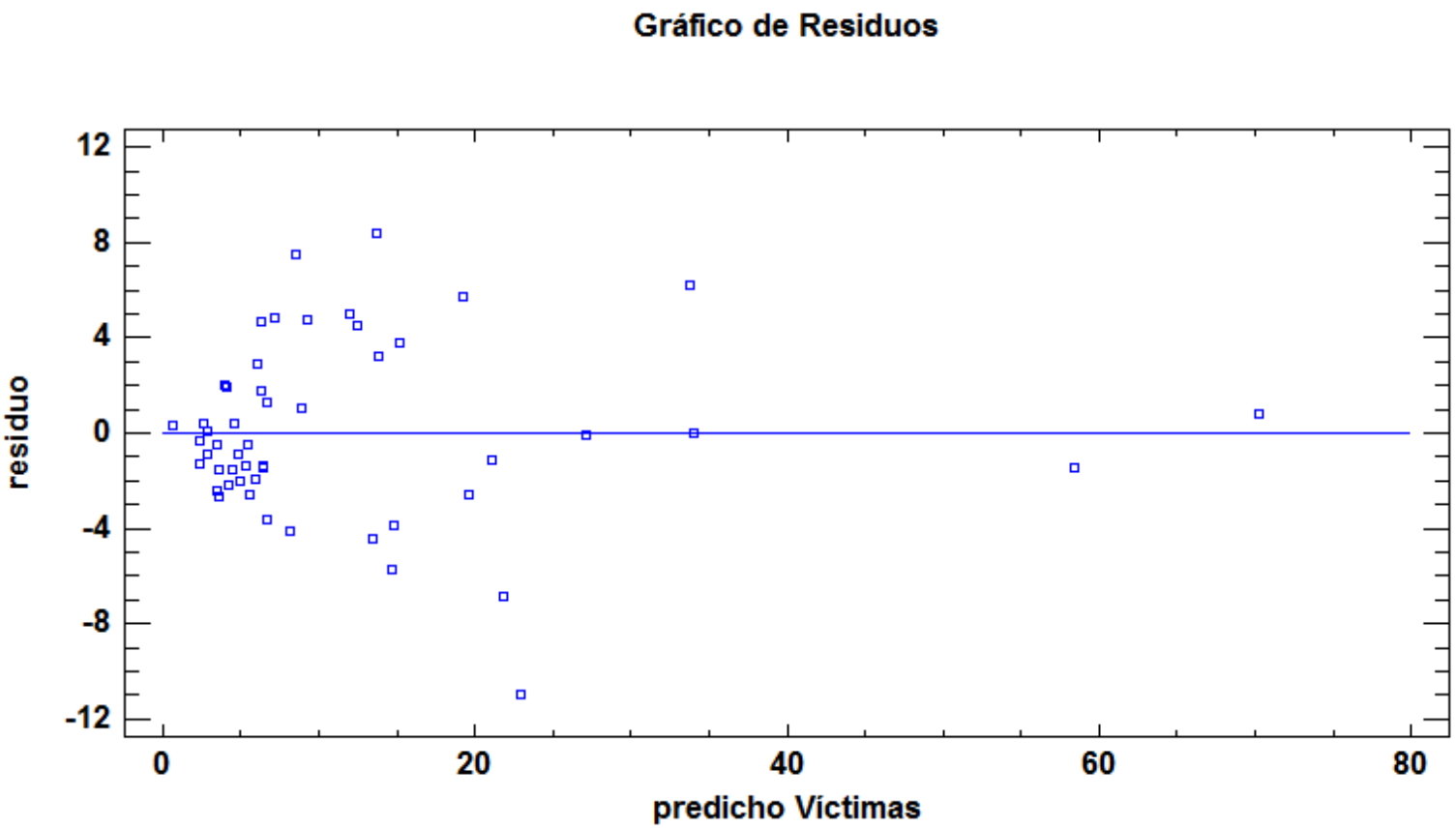
Denuncias
Llamadas al 016
Nº matrimonios católicos
Nº matrimonios de otra religión
Nº matrimonios civiles
Órdenes de protección
Población femenina
Población total

1) Análisis exploratorio


Denuncias
Llamadas 016
Matr. católicos
Matr. civiles
Matr. otra religión
Órdenes protección
Población femenina
Población total
Victimas

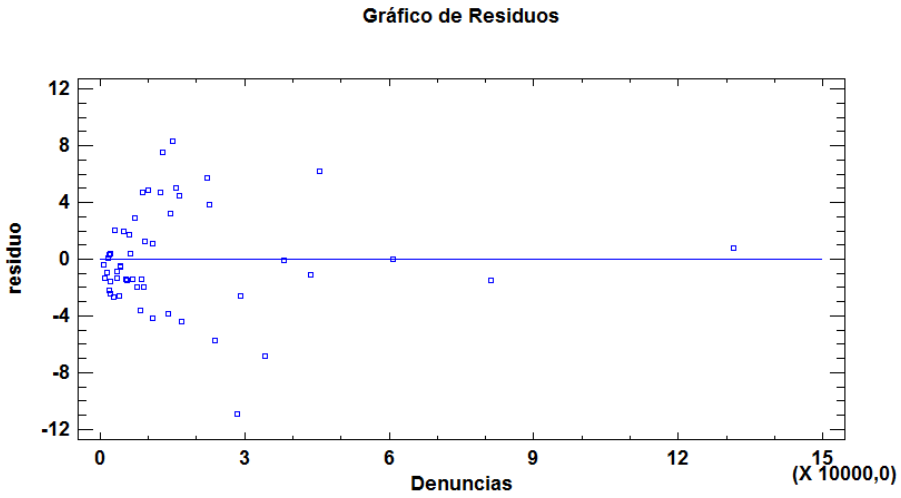
2) Regresión múltiple inicial

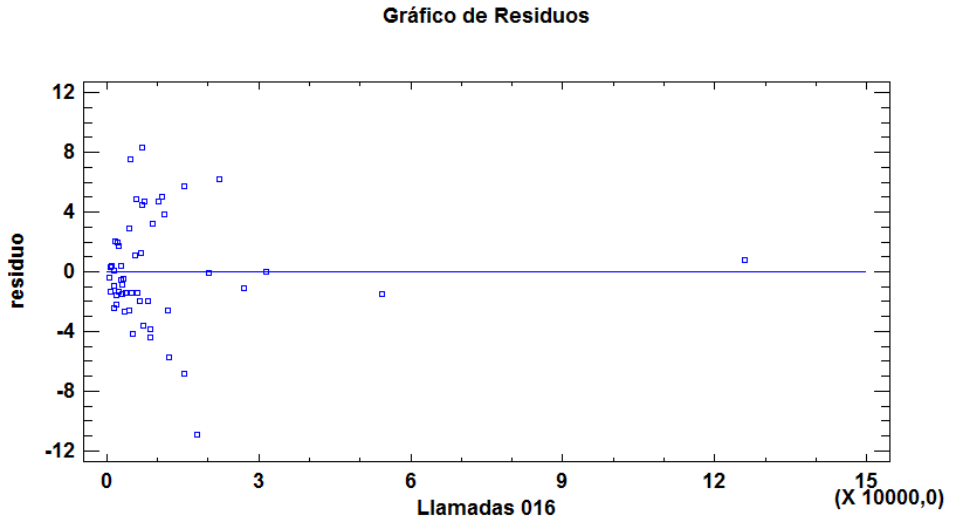
El gráfico de residuos frente a valores predichos nos permite evaluar la LINEALIDAD

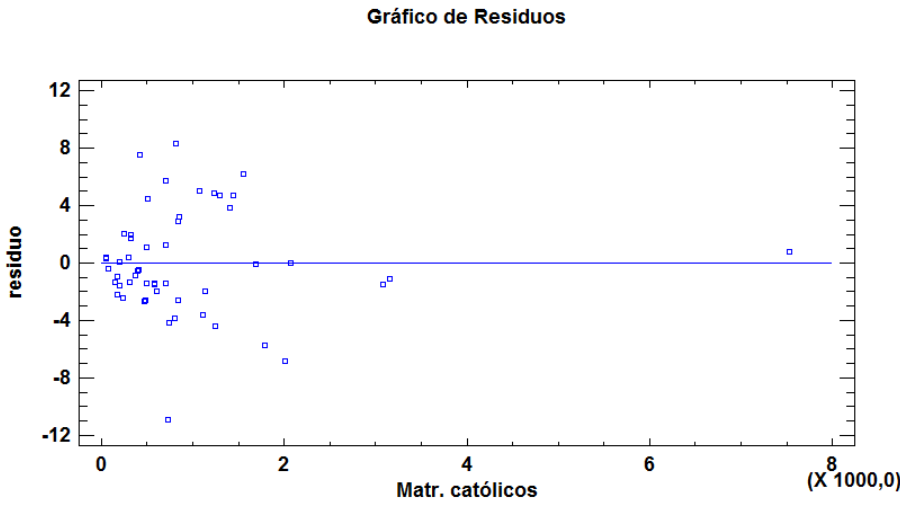


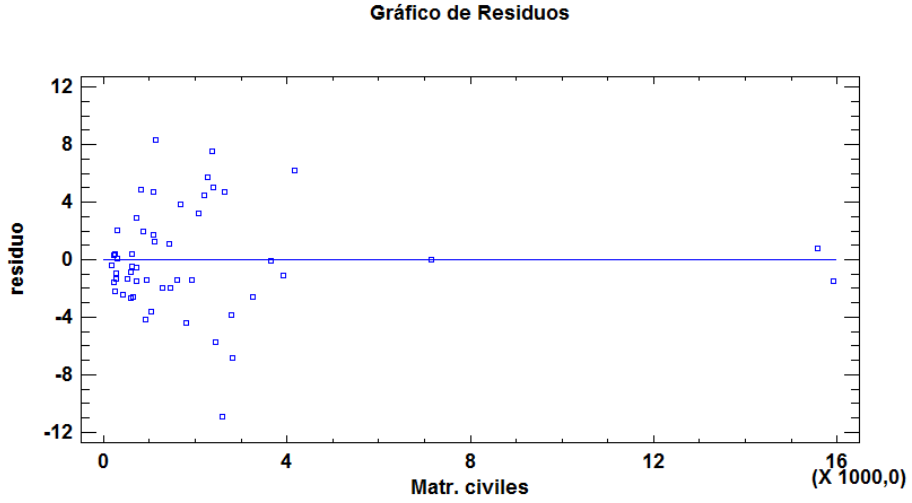
Distribución no aleatoria → mal ajuste lineal

Los gráficos de residuos frente a valores de las 8 variables muestran un comportamiento similar:

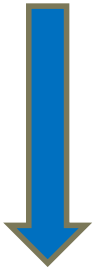








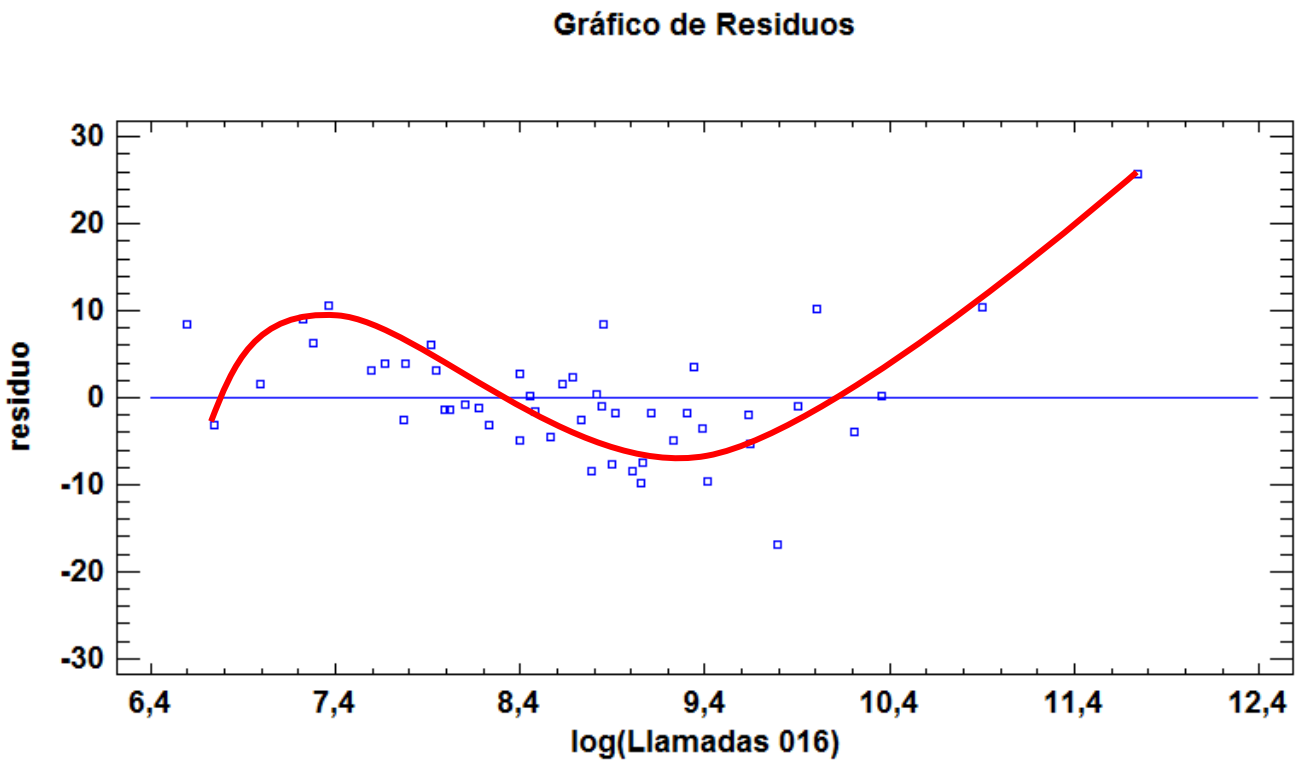
Los datos no tienen un comportamiento lineal y no son homocedásticos.



TRANSFORMACIÓN

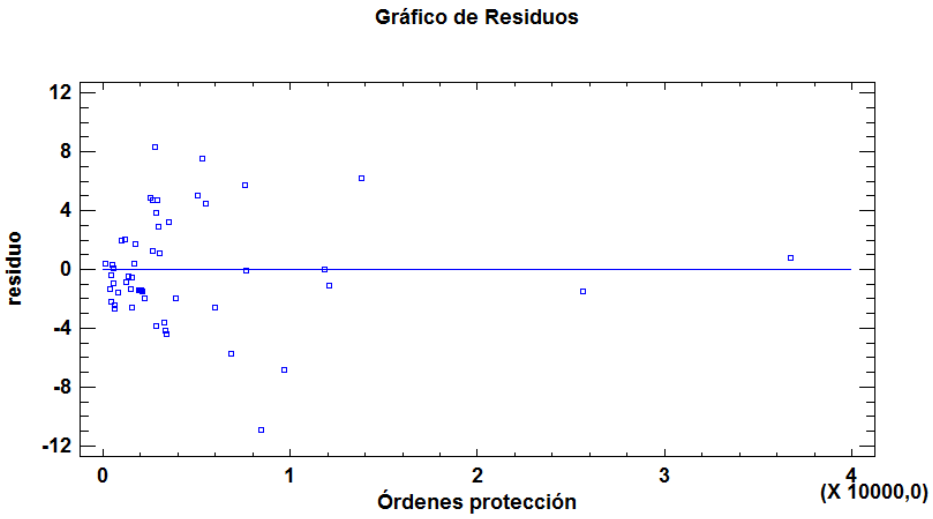
Probamos con el logaritmo neperiano de las las variables independientes.

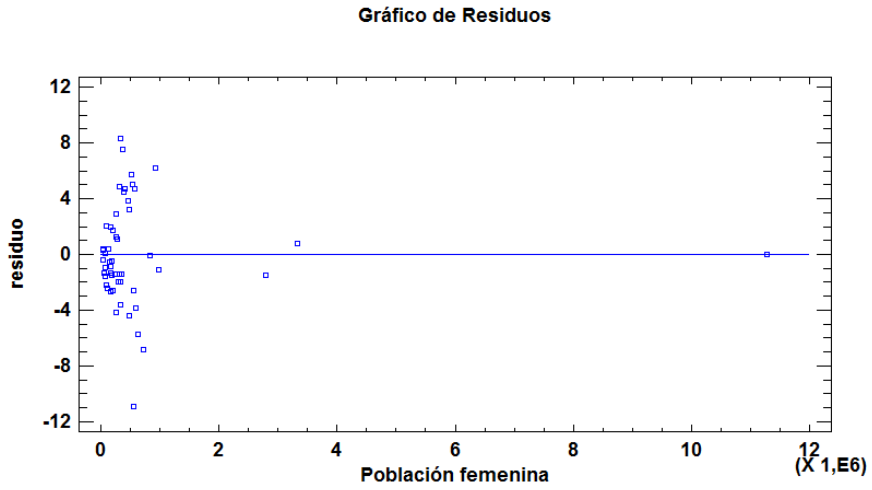
En todas aparecen patrones curvos más o menos acusados.



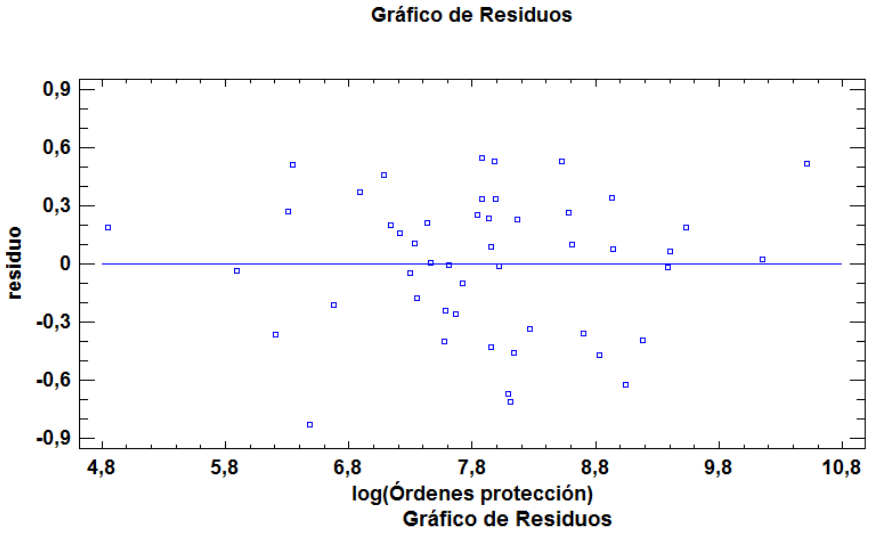
Transformamos también la variable respuesta.

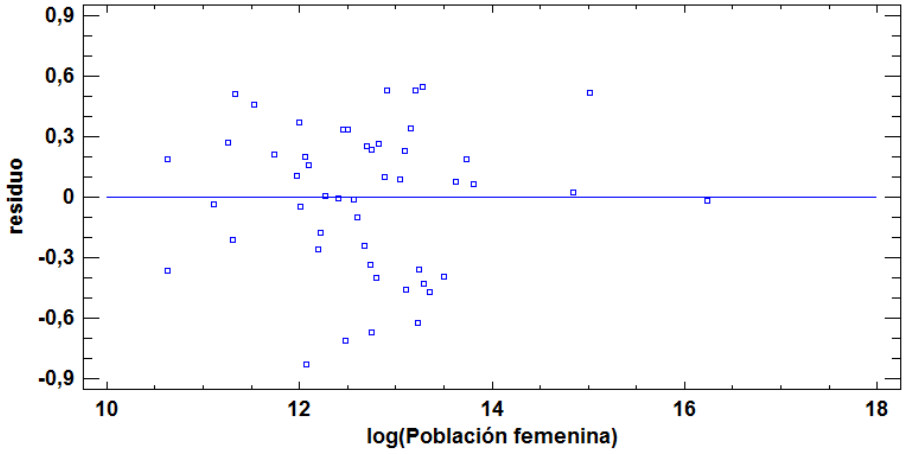
Antes





Después

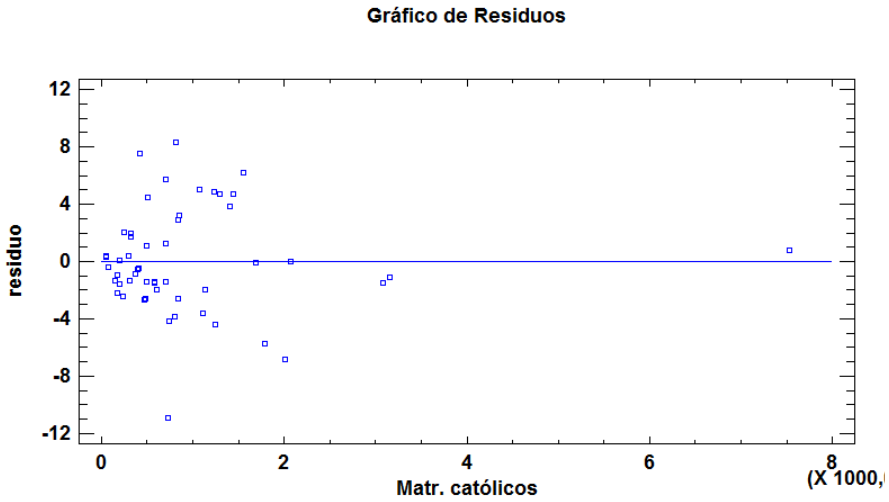


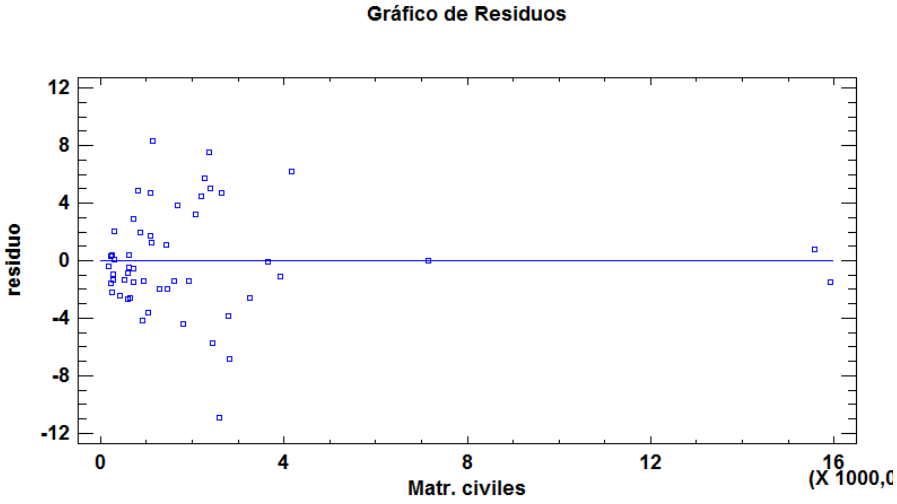


Así hemos conseguido linealidad y homocedasticidad

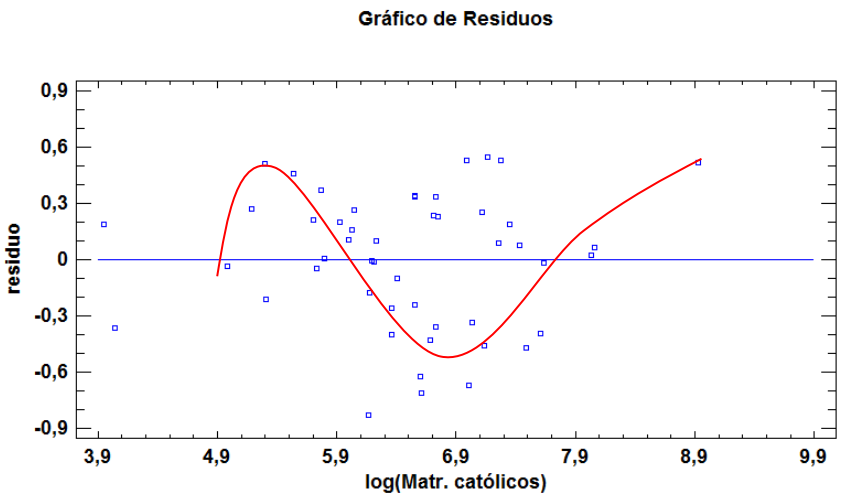
Sin embargo, no parece una buena transformación para el número de matrimonios.

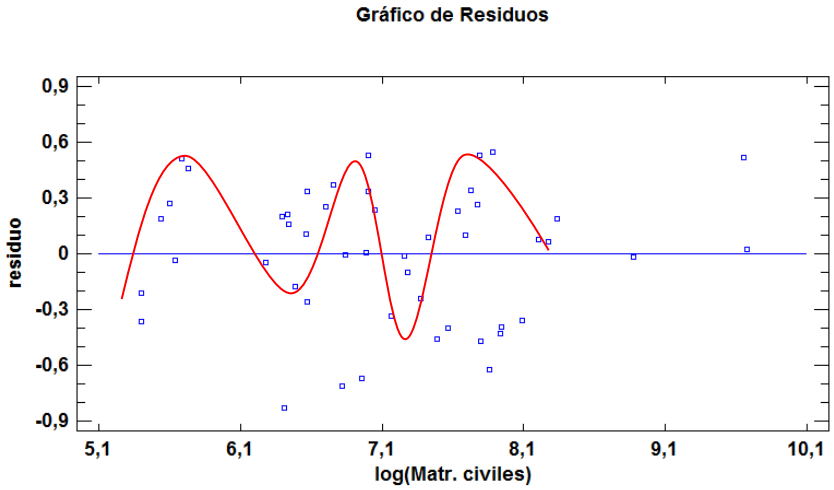
Antes





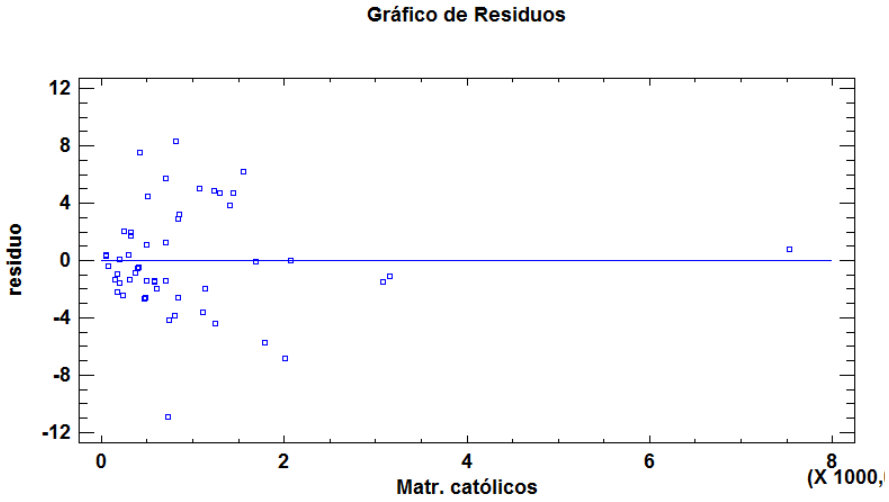
Después

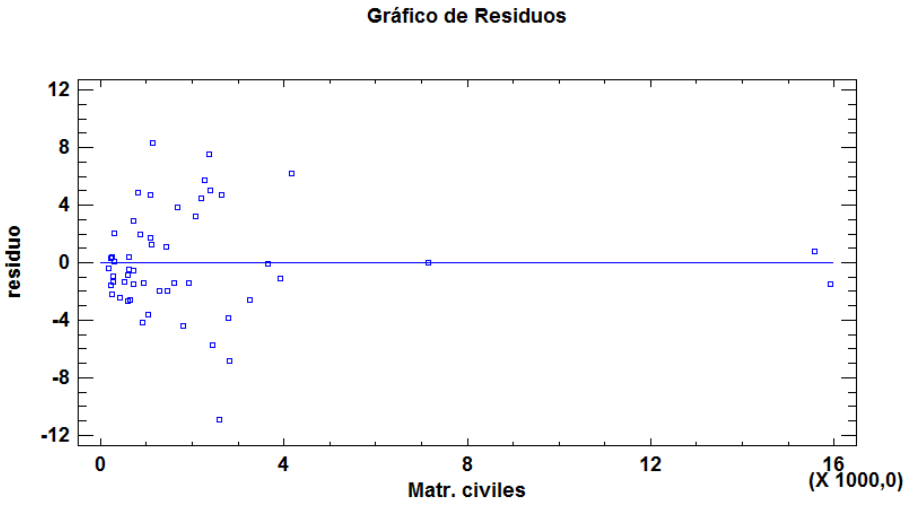




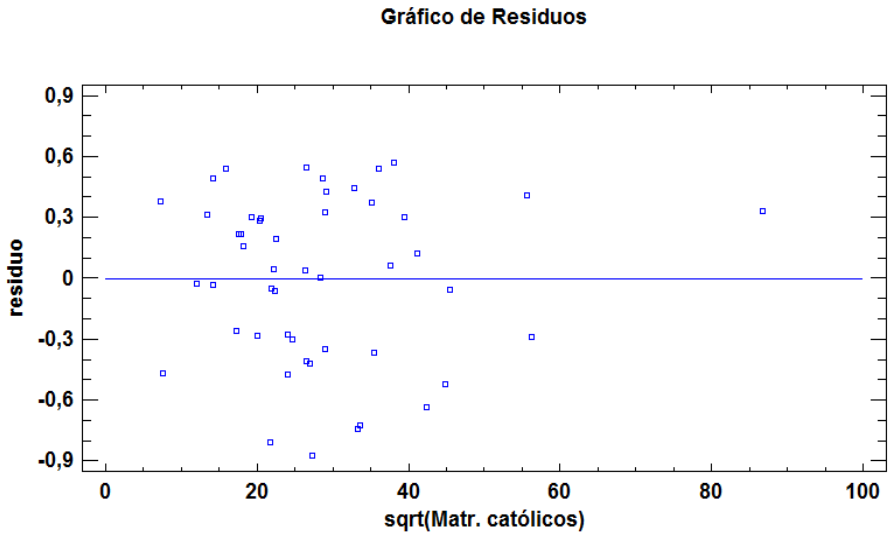
Con la transformación raíz cuadrada:

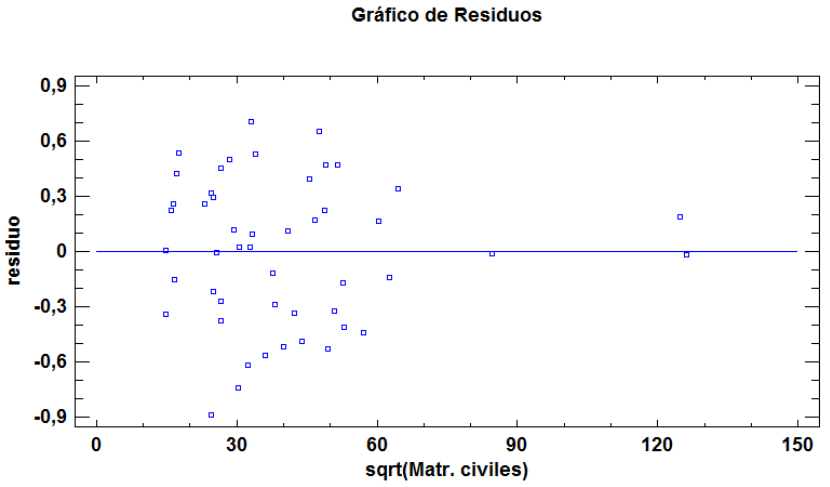
Antes





Después





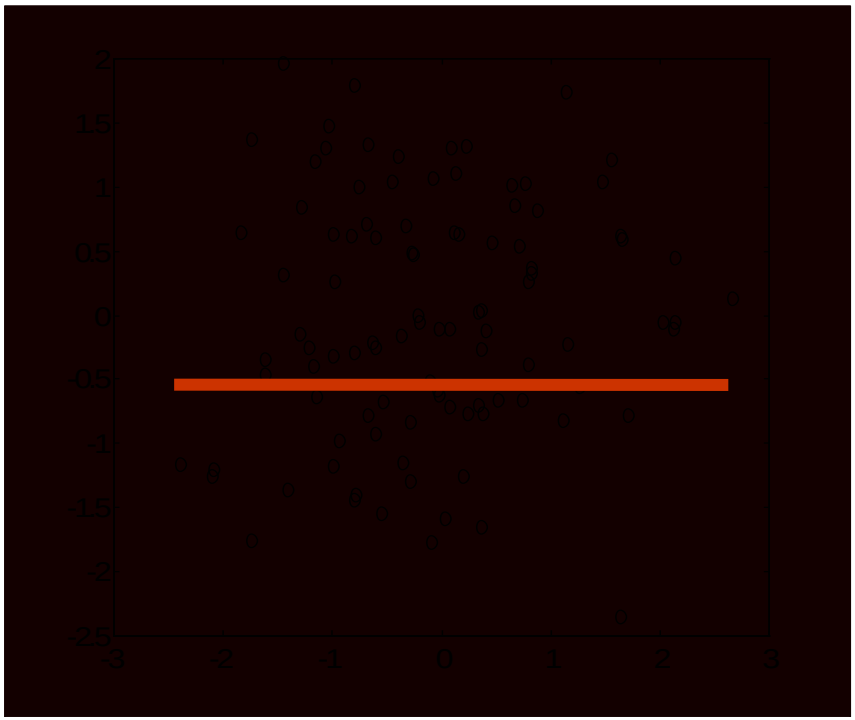
3) Análisis de significatividad

Variable significativa
es aquella que aporta información sobre Y no incluida en el resto de las variables.

Por tanto,
será relevante incluirla en la regresión.

Una variable será no significativa si:

$$\beta_i = 0$$



Hacemos el contraste:

H_0: $\beta_i = 0$

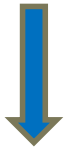
H_1: $\beta_i \neq 0$

Parámetro	Estimación	*Error Est,*	*Estadístico T*	*Valor-P*
CONSTANTE	-6,3866	2,94291	-2,17017	0,0360
log(Denuncias)	1,39519	0,303231	4,60107	0,0000
log(Llamadas 016)	-0,576818	0,328657	-1,75508	0,0869
sqrt(Matr. católicos)	-0,00516775	0,0117402	-0,440176	0,6622
sqrt(Matr. civiles)	0,010688	0,00754925	1,41576	0,1646
log(Matr. otra religión)	-0,139823	0,0863812	-1,61867	0,1134
log(Órdenes protección)	-0,142975	0,206453	-0,69253	0,4926
log(Población femenina)	-0,0879434	0,211083	-0,41663	0,6792
log(Población total)	0,217323	0,437677	0,496538	0,6222

R^2 ajustado = 82,1094 %

$$p > 0.05$$

Aceptamos Ho ⇨ variable no significativa.



Parámetro	*Estimación*	*Error est.*	*Estadístico T*	*Valor-P*
CONSTANTE	-6,29401	2,90478	-2,16678	0,0361
log(Denuncias)	1,38636	0,299426	4,63007	0,0000
log(Llamadas 016)	-0,581994	0,325095	-1,79023	0,0808
sqrt(Matr. católicos)	-0,00463361	0,0115518	-0,401117	0,6904
sqrt(Matr. civiles)	0,0102075	0,00738509	1,38218	0,1744
log(Matr. otra religión)	-0,139742	0,085506	-1,63429	0,1099
log(Órdenes protección)	-0,133856	0,20321	-0,658706	0,5138
log(Población total)	0,131003	0,381627	0,343276	0,7331

R^2 ajustado = 82,47 %

Parámetro	*Estimación*	*Estimación*	*Error Est.*	*Estadístico T*	*Valor-P*
CONSTANTE		-5,05848	0,930079	-5,43876	0,0000
log(Denuncias)		1,35785	0,232888	5,83047	0,0000
log(Llamadas 016)		-0,630103	0,246353	-2,55772	0,0141
sqrt(Matr. civiles)		0,0105842	0,00627977	1,68544	0,0990
log(Matr. otra religión)		-0,146827	0,0752136	-1,95214	0,0573

R^2 ajustado = 83,43 %

log(Víctimas) = -5,05848 + 1,35785*log(Denuncias) - 0,630103*log(Llamadas 016) + 0,0105842*sqrt(Matr. civiles) - 0,146827*log(Matr. otra religión)

R^2 ajustado = 83,43 %

4) Diagnosis del modelo

Linealidad

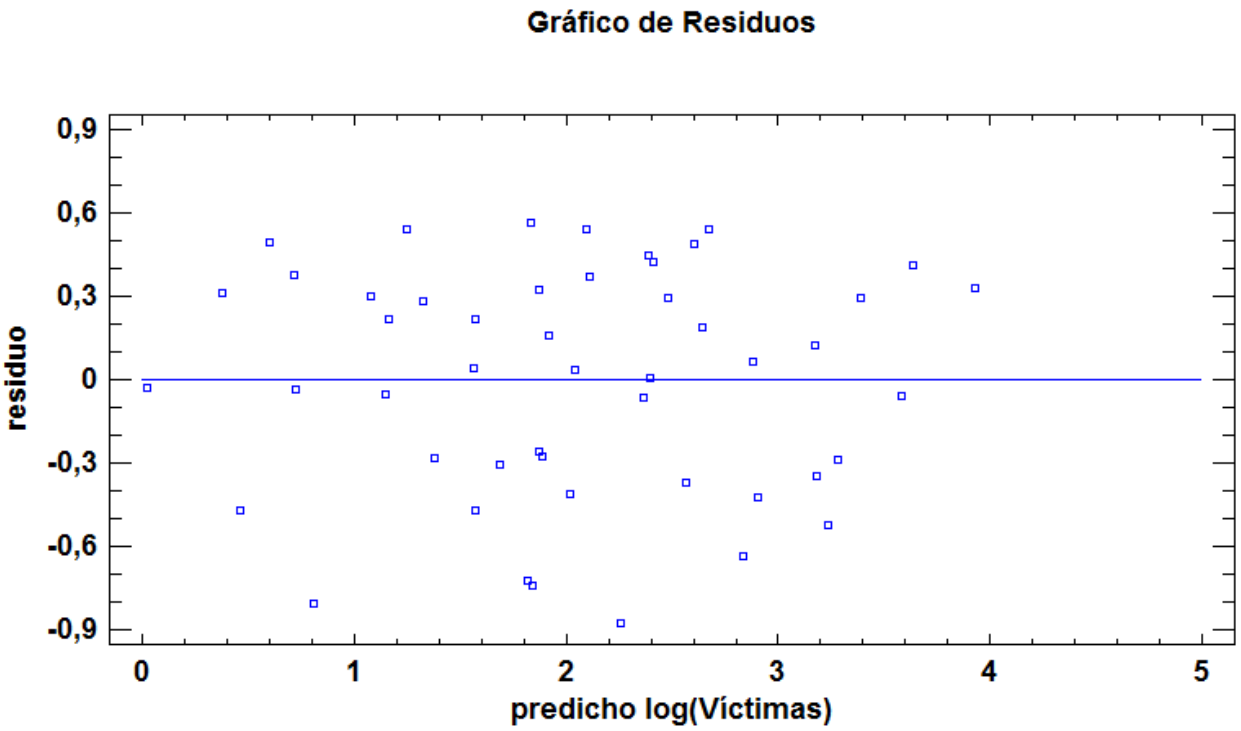
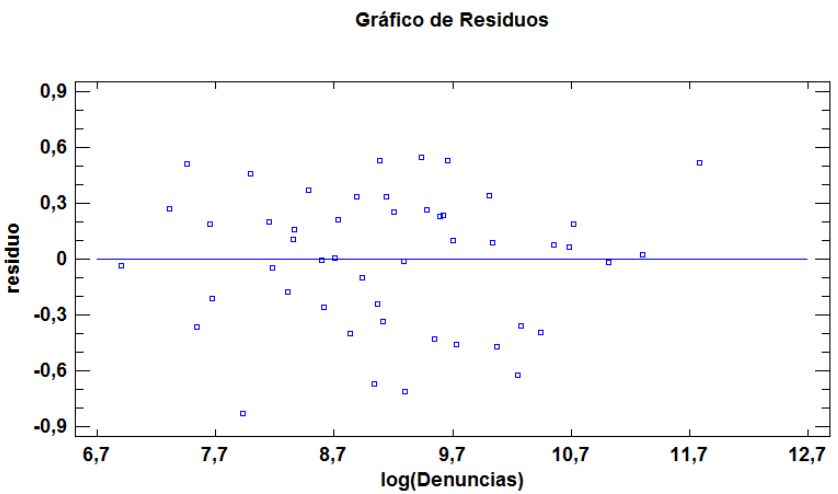
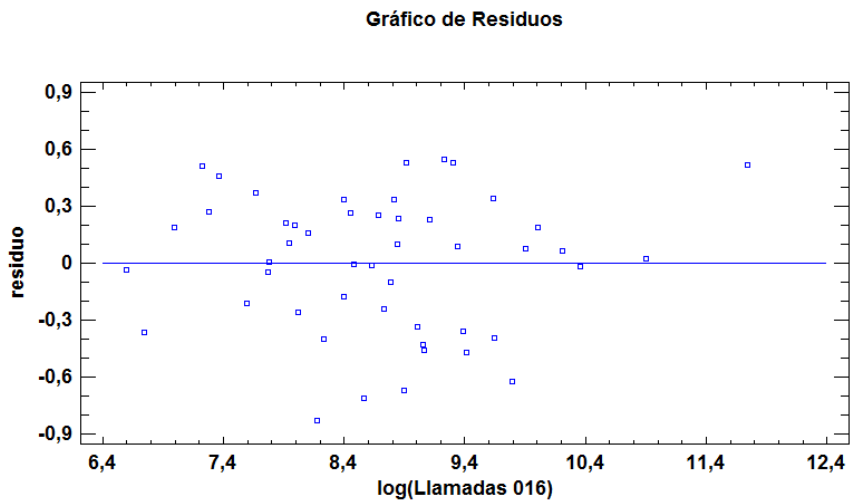
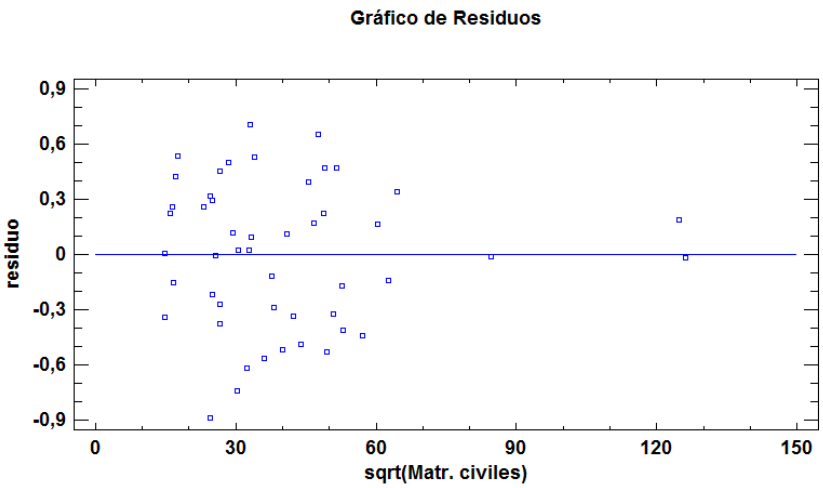


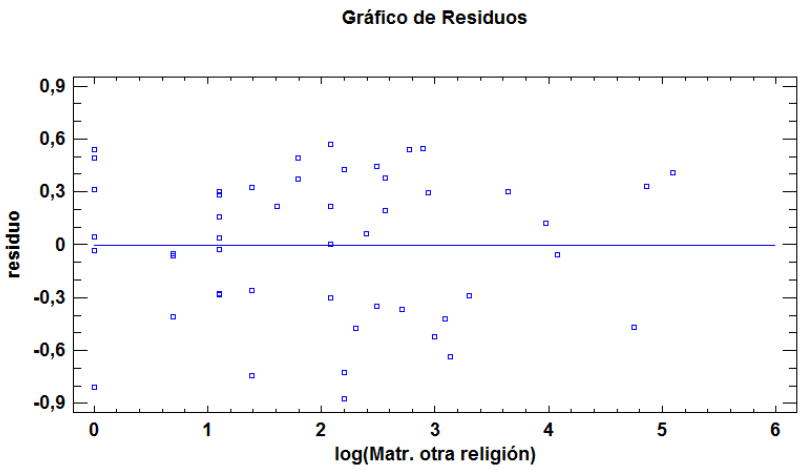
Gráfico de Residuos
residuo
0,9
0,6
0,3
0
-0,3
-0,6
-0,9
0
1
2
3
4
5
predicho log(Víctimas)

Homocedasticidad





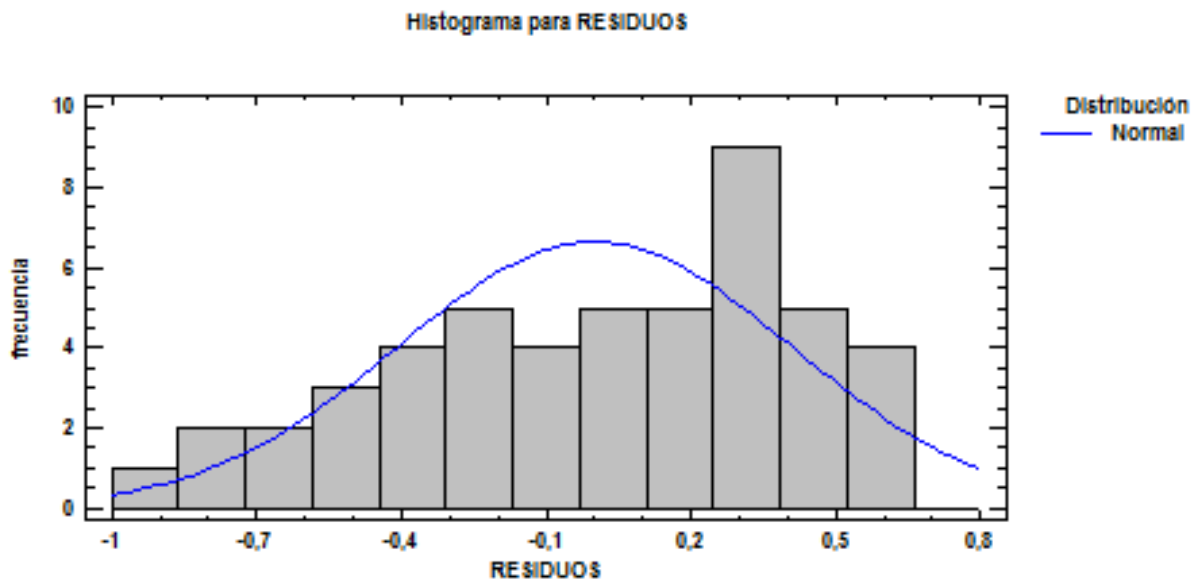




- Independencia

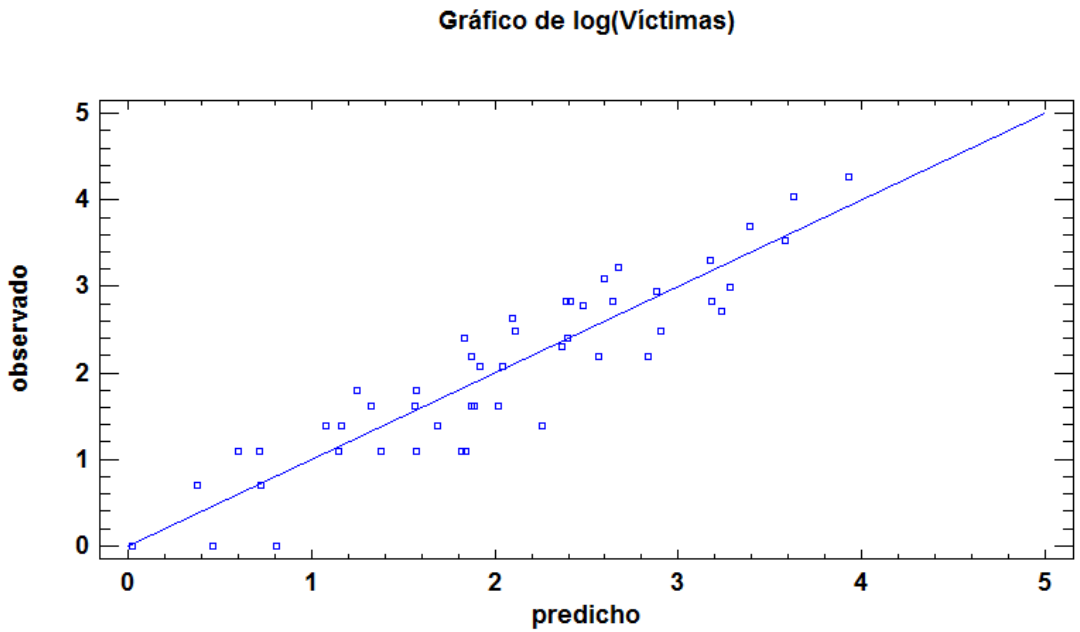
Estadístico Durbin-Watson = 2,40431
(p-valor=0,9045)

- Normalidad



Chi-Cuadrada = 9,72521 con 8 g.l.

p-valor = 0,284845



log(Víctimas) = -5,05848
+1,35785*log(Denuncias)
-0,630103*log(Llamadas 016)
+0,0105842*sqrt(Matr. civiles)
-0,146827*log(Matr. otra religión)

Algunas fuentes de interés...

Blalock, H. M. (1966). *Estadística social*. Fondo de cultura económica.

Instituto Nacional de Estadística. www.ine.es

Peña, D. (2002). *Regresión y diseño de experimentos*. Alianza editorial.

Portal Estadístico. Delegación del Gobierno para la Violencia de Género. http://estadisticasviolenciagenero.msssi.gob.es/