Análisis de calidad del agua de ríos neoyorquinos

Enfocado en la regresión lineal

Manuel Verdejo García

27/02/2025

Índice

1. **Antecedentes**: se ha realizado un estudio medioambiental en veinte ríos neoyorquinos.
2. **Objetivos:**
   2.1. Predecir la calidad del agua en función de las variables Xi, sin necesidad de medir directamente la calidad del agua.
   2.2. Identificar cuáles de las cuatro variables independientes tienen una influencia significativa en la calidad del agua.
   2.3. Determinar el mejor subconjunto de regresión, es decir, el conjunto óptimo de variables que mejor explican la calidad del agua.
3. **Variables:**

Se consideran cinco variables cuantitativas continuas: una variable dependiente (Y) y cuatro variables independientes (X_1, X_2, X_3, X_4). A continuación, se describen brevemente:

- Y: Concentración de nitrógeno (calidad del agua).
- X_1: Porcentaje de terrenos agrícolas adyacentes al río.
- X_2 :Porcentaje de terrenos forestales adyacentes al río.
- X_3: Porcentaje de terrenos residuales adyacentes al río.
- X_4: Porcentaje de terrenos comerciales e industriales adyacentes al río.

4. **Modelo:**

Dado el tipo de variables consideradas, se emplea un modelo de regresión lineal múltiple, que permite analizar la influencia de las variables independientes en la variable dependiente. El modelo a utilizar es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Para que el modelo sea válido, deben cumplirse los siguientes supuestos:

- La esperanza matemática de los residuos es igual a cero.
- La varianza del error es constante e independiente de X (homocedasticidad).
- Los errores son independientes entre sí, es decir:

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$$

- Los errores siguen una distribución normal.
- Todas las observaciones tienen igual importancia en la estimación de los parámetros.

5. **Estudio Piloto:**

No procede en este caso, ya que se supone que se ha realizado un estudio previo.

6. Estudio de la Forma y Tamaño de la Muestra:

De igual manera, no aplica en este contexto.

7. Análisis de Datos:

La muestra está compuesta por 20 ríos, mientras que la población objetivo corresponde a todos los ríos de Nueva York.

7.1. Estadística descriptiva: no procede.

7.2. Análisis Estadístico:

7.2.1. Bloque A: Modelización y Estimación de los Coeficientes de Regresión. Modelo Estimado.

- Diagrama de dispersión:

Correlaciones

		CONCENTR	AGRICOLA	FORESTAL	COMIND	RESIDENCIAL
Correlación de Pearson	CONCENTR	1,000	,401	-,773	,532	,566
	AGRICOLA	,401	1,000	-,683	-,346	-,242
	FORESTAL	-,773	-,683	1,000	-,309	-,503
	COMIND	,532	-,346	-,309	1,000	,859
	RESIDENCIAL	,566	-,242	-,503	,859	1,000
Sig. (unilateral)	CONCENTR	.	,040	<,001	,008	,005
	AGRICOLA	,040	.	,000	,068	,152
	FORESTAL	,000	,000	.	,093	,012
	COMIND	,008	,068	,093	.	,000
	RESIDENCIAL	,005	,152	,012	,000	.
N	CONCENTR	20	20	20	20	20
	AGRICOLA	20	20	20	20	20
	FORESTAL	20	20	20	20	20
	COMIND	20	20	20	20	20
	RESIDENCIAL	20	20	20	20	20

Esto no vale.

- Tabla ANOVA:

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,570	4	,642	9,154	<,001[b]
	Residuo	1,053	15	,070		
	Total	3,623	19			

a. Variable dependiente: CONCENTR

b. Predictores: (Constante), RESIDENCIAL, AGRICOLA, COMIND, FORESTAL

Como la significación es menor a 0.01 podemos rechazar que las todos los coeficientes sean cero. Falta diagnosticar la normalidad de Y.

- Tabla bondad de ajuste:

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Cambio en R cuadrado
1	,842[a]	,709	,632	,26492	,709

a. Predictores: (Constante), RESIDENCIAL, AGRICOLA, COMIND, FORESTAL

b. Variable dependiente: CONCENTR

A partir del R cuadrado, podemos decir que las variables independientes explican un 71% aproximadamente de la variable dependiente.

- Tabla de coeficientes y tolerancia:

Coeficientes[a]

	Coeficientes no estandarizados		Coeficientes estandarizados			95,0% intervalo de confianza para B		Correlaciones			Estadísticas de colinealidad	
	B	Desv. Error	Beta	t	Sig.	Límite inferior	Límite superior	Orden cero	Parcial	Parte	Tolerancia	VIF
(Constante)	1,722	1,234		1,396	,183	-,908	4,353					
AGRICOLA	,006	,015	,196	,386	,705	-,026	,038	,401	,099	,054	,075	13,277
FORESTAL	-,013	,014	-,530	-,931	,367	-,043	,017	-,773	-,234	-,130	,060	16,727
COMIND	,305	,164	,528	1,862	,082	-,044	,654	,532	,433	,259	,241	4,145
RESIDENCIAL	-,007	,034	-,106	-,214	,834	-,079	,065	,566	-,055	-,030	,079	12,682

Vamos a estimar el modelo:

$$CNTR = \beta_0 + \beta_1 X_1 + \cdots + \beta_4 X_4$$

Cada fila de la tabla representa un β. Si $\beta_0 = 0$, se seleccionan los coeficientes de la columna Beta; si es distinto de 0, como es el caso, se toman los coeficientes de la columna B.

Para ello, se realiza un test de hipótesis, cuya significación se encuentra en la columna Sig. Si este valor es mayor a 0.05, no existen evidencias suficientes para rechazar que sea igual a 0. Como es el caso para todas las variables no podemos rechazarlo para ninguna. Analizando la significancia de todos los coeficientes determinamos cuáles son influyentes y cuáles no, siendo los más influyentes en este modelo “Agrícola” y “Resid”.

Si una variable está altamente correlacionada con las demás, puede causar problemas, por lo que buscamos una tolerancia cercana a 1. Una baja tolerancia indica un problema de colinealidad. Como la tolerancia de todas las variables es baja vamos a buscar el mejor subconjunto de regresión, es decir, intentaremos sacar las peores variables del modelo.

Utilizamos el método de Forward Selection, que agrega las variables una por una, seleccionando las más significativas hasta que encuentra una que no lo sea.

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	FORESTAL	.	Avanzar (Criterio: Probabilidad-de-F-para-entrar <= ,050)
2	COMIND	.	Avanzar (Criterio: Probabilidad-de-F-para-entrar <= ,050)

a. Variable dependiente: CONCENTR

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2,167	1	2,167	26,799	<,001[b]
	Residuo	1,456	18	,081		
	Total	3,623	19			
2	Regresión	2,512	2	1,256	19,236	<,001[c]
	Residuo	1,110	17	,065		
	Total	3,623	19			

a. Variable dependiente: CONCENTR

b. Predictores: (Constante), FORESTAL

c. Predictores: (Constante), FORESTAL, COMIND

La primera tabla muestra las variables que componen el subconjunto de regresión, Forestal y Comind. En la segunda tabla, cada fila representa el modelo en cada paso de adición de variables, por lo que nos interesa la última fila (correspondiente al mejor subconjunto elegido por el programa).

Resumen del modelo[c]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Estadísticos de cambio					Criterio de información de Akaike
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. Cambio en F	
1	,773[a]	,598	,576	,28437	,598	26,799	1	18	<,001	-48,407
2	,833[b]	,694	,657	,25555	,095	5,288	1	17	,034	-51,824

a. Predictores: (Constante), FORESTAL

b. Predictores: (Constante), FORESTAL, COMIND

c. Variable dependiente: CONCENTR

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			95,0% intervalo de confianza para B		Correlaciones			Estadísticas de colinealidad	
		B	Desv. Error	Beta	t	Sig.	Límite inferior	Límite superior	Orden cero	Parcial	Parte	Tolerancia	VIF
1	(Constante)	2,347	,238		9,844	<,001	1,846	2,848					
	FORESTAL	-,019	,004	-,773	-5,177	<,001	-,027	-,011	-,773	-,773	-,773	1,000	1,000
2	(Constante)	2,096	,240		8,718	<,001	1,589	2,604					
	FORESTAL	-,016	,003	-,673	-4,769	<,001	-,024	-,009	-,773	-,756	-,640	,905	1,105
	COMIND	,188	,082	,325	2,300	,034	,015	,360	,532	,487	,309	,905	1,105

a. Variable dependiente: CONCENTR

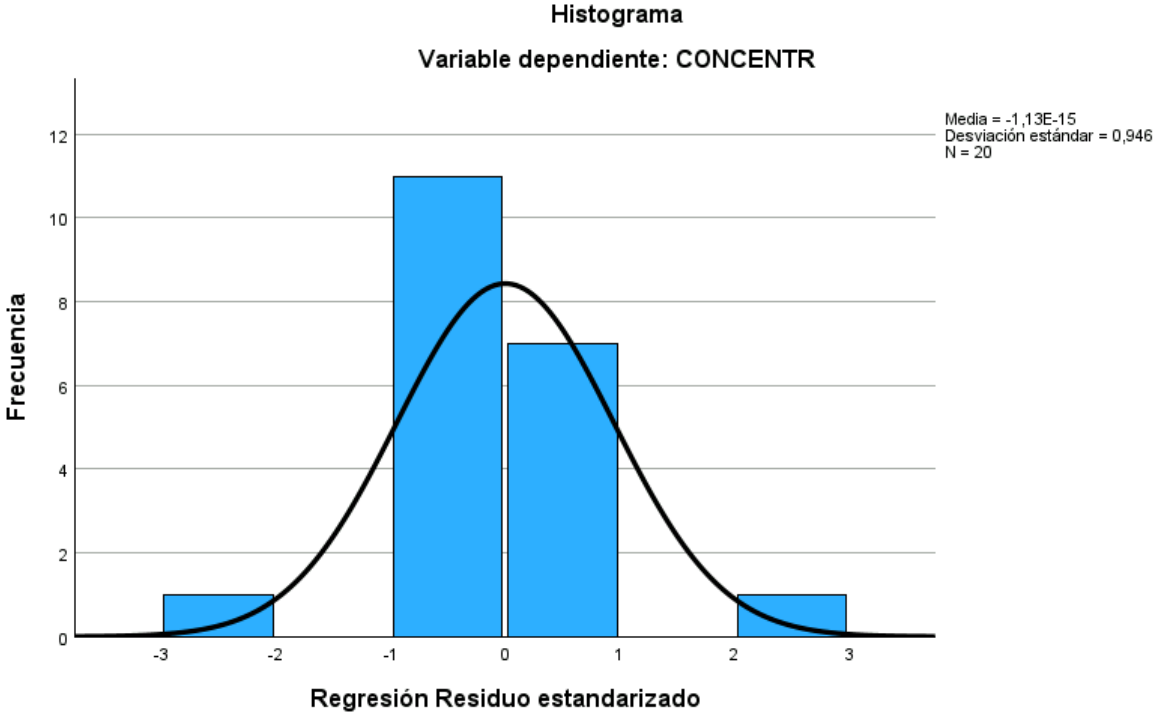
En la primera tabla observamos cómo mejora el R cuadrado ajustado en el segundo subconjunto. A partir de los coeficientes B de la segunda tabla, se construye el modelo:

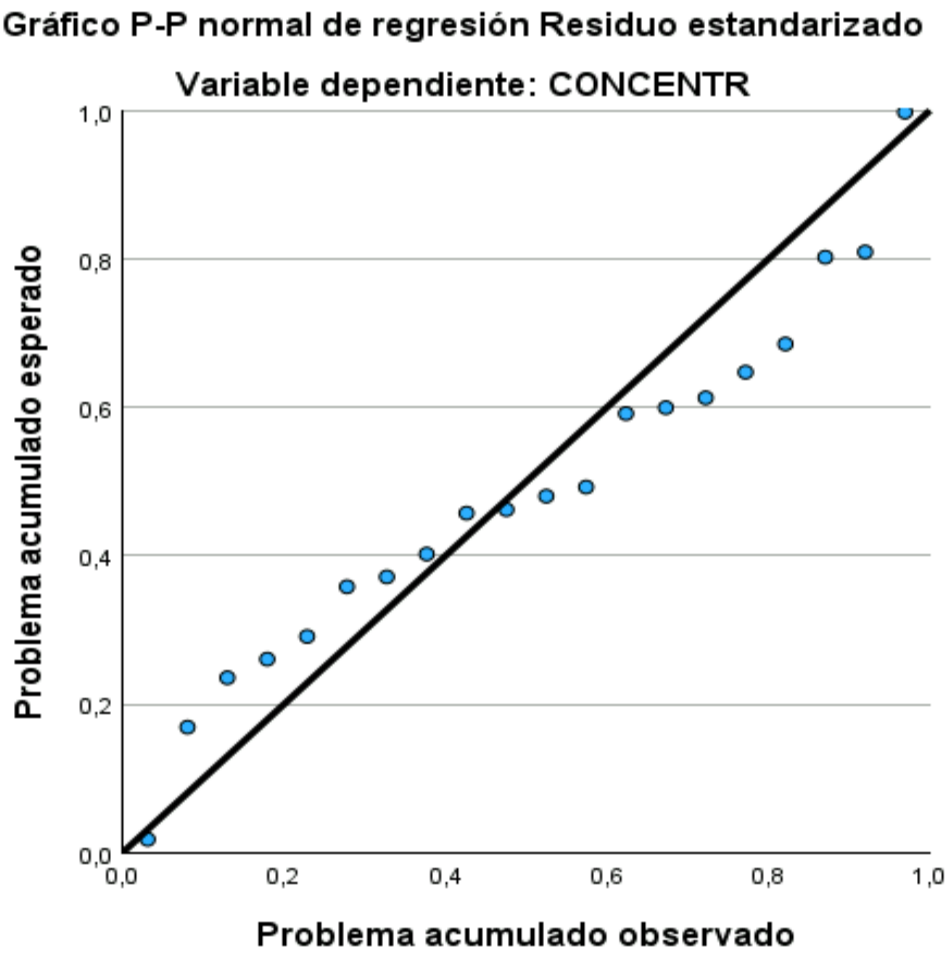
$$Y = B_0 + B_1 X_1 + \cdots + B_n X_n$$

Como la significación de la variable está entre 0.01 y 0.05 (0.034), es dudosa, sin embargo indicamos que es menor a 0.05. Además, analizamos la tolerancia del modelo que es relativamente cercana a 1 (0.905) por lo que no podemos decir que sea mala.

7.2.2. Bloque B: Diagnóstico de las Hipótesis asociadas al Modelo

- Procedemos ahora a analizar la normalidad de la variable dependiente (concentración).





En estas dos representaciones gráficas buscamos que en la primera su distribución se parezca a la normal y en la segunda que se aproxime a la recta. Como podemos ver, ninguno de los dos casos se da. Procedemos a hacer el test de Kolmogorov-Smirnov.

Prueba de Kolmogorov-Smirnov para una muestra

			Standardized Residual
N			20
Parámetros normales[a,b]	Media		,0000000
	Desv. estándar		,88852332
Máximas diferencias extremas	Absoluta		,142
	Positivo		,142
	Negativo		-,109
Estadístico de prueba			,142
Sig. asin. (bilateral)[c]			,200[d]
Sig. Monte Carlo (bilateral)[e]	Sig.		,346
	Intervalo de confianza al 99%	Límite inferior	,334
		Límite superior	,358

a. La distribución de prueba es normal.

b. Se calcula a partir de datos.

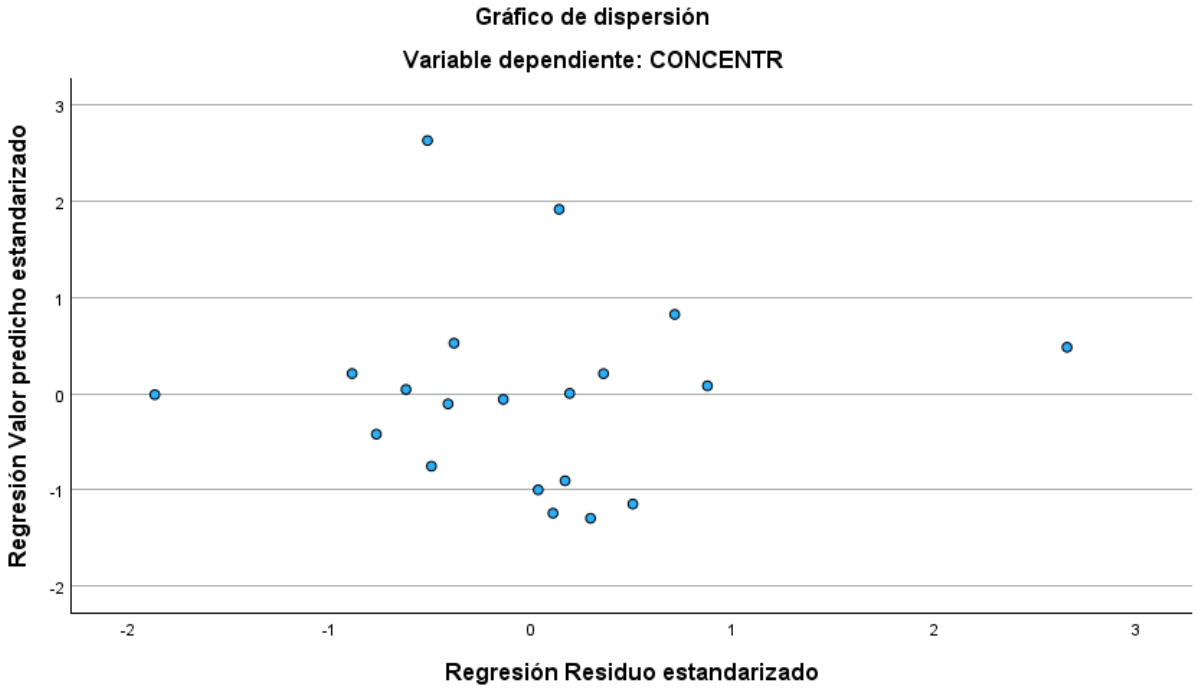
c. Corrección de significación de Lilliefors.

d. Esto es un límite inferior de la significación verdadera.

e. El método de Lilliefors basado en las muestras 10000 Monte Carlo con la semilla de inicio 299883525.

Planteamos el test de hipótesis nula X normal e hipótesis alternativa su opuesto. Como podemos ver, su significancia (0.2) es mayor que 0.05 por lo que no podemos rechazar que la distribución sea normal.

- Veamos ahora si es homocedástica:



Como podemos observar hay dos o más valores que se escapan de una posible banda de anchura constante, por lo tanto no podemos afirmar que lo sea y deberíamos hacer unas transformaciones.

- Veamos ahora si los errores son independientes:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	,842[a]	,709	,632	,26492	1,507

a. Predictores: (Constante), RESIDENCIAL, AGRICOLA, COMIND, FORESTAL

b. Variable dependiente: CONCENTR

Como el coeficiente de Durbin-Watson (1.507) está comprendido entre 1,5 y 2,5 no podemos rechazar su independencia.

- La multicolinealidad ha sido estudiada en los Métodos Paso a Paso construyendo el Mejor Modelo de Regresión con la tolerancia.

7.3. Bloque C: Diagnóstico de Observaciones Anómalas

	CONCENTR	AGRICOL..	FORREST...	COMIND	RESIDENCIAL	ZRE_1	ZRE_2	PRE_1	ZRE_3	SDR_1	MAH_1	COO_1	LEV_1
1	1,10	26,00	63,00	,29	1,20	-,13612	-,13612	1,13606	-,13612	-,13917	1,06382	,00049	,05599
2	1,01	29,00	57,00	,09	,70	-,61869	-,61869	1,17390	-,61869	-,63314	,63203	,00758	,03326
3	1,90	54,00	26,00	,58	1,80	,14099	,14099	1,86265	,14099	,18250	7,44160	,00563	,39166
4	1,00	2,00	84,00	1,98	1,90	-,88613	-,88613	1,23475	-,88613	-3,79574	16,08888	13,21955	,84678
5	1,99	3,00	27,00	3,11	29,40	-,51215	-,51215	2,12568	-,51215	-4,84434	17,51956	65,42632	,92208
6	1,42	19,00	61,00	,56	3,40	,87653	,87653	1,18779	,87653	,89475	,05833	,00909	,00307
7	2,04	16,00	60,00	1,11	5,60	2,66045	2,66045	1,33520	2,66045	3,90560	,85597	,16430	,04505
8	1,65	40,00	43,00	,24	1,30	,71429	,71429	1,46077	,71429	,76806	2,06641	,02289	,10876
9	1,01	28,00	62,00	,15	1,10	-,41018	-,41018	1,11866	-,41018	-,43104	1,77803	,00659	,09358
10	1,21	26,00	60,00	,23	,90	,19315	,19315	1,15883	,19315	,19315	,26816	,00055	,01411
11	1,33	26,00	53,00	,18	,90	,36104	,36104	1,23435	,36104	,38862	2,57931	,00730	,13575
12	,75	15,00	75,00	,16	,70	-,49262	-,49262	,88050	-,49262	-,51064	1,23845	,00714	,06518
13	,73	6,00	84,00	,12	,50	,11039	,11039	,70076	,11039	,11691	2,22520	,00059	,11712
14	,80	3,00	81,00	,35	,80	,03692	,03692	,79022	,03692	,03881	2,00383	,00006	,10546
15	,76	2,00	89,00	,35	,70	,29673	,29673	,68139	,29673	,32095	2,78104	,00535	,14637
16	,87	6,00	82,00	,15	,50	,50641	,50641	,73584	,50641	,53238	1,67993	,00956	,08842
17	,80	22,00	70,00	,22	,90	-,76576	-,76576	1,00286	-,76576	-,83100	2,24907	,02855	,11837
18	,87	4,00	75,00	,18	,40	,17034	,17034	,82487	,17034	,20237	5,44880	,00444	,28678

Analizamos ahora si hay valores atípicos. Como los más problemáticos son las Observaciones influyentes, nos fijamos en los valores más altos de la columna de la distancia de Cook. Como se puede ver, las observaciones 4 y 5 tienen unos valores muy altos en comparación al resto. Para solucionar este problema podemos rectificar errores en la toma de datos, eliminar observaciones, usar estimadores alternativos (trabajar con la mediana en vez de la media por ejemplo), hacer un modelo no lineal, recoger más datos y hacer el procedimiento de mínimos cuadrados con robustez.