UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Instituto del Cálculo

Trabajo final integrador

Especialización en Estadística Matemática

ESTIMACIÓN DE LA SOLUBILIDAD EN AGUA DE ALCOHOLES ALIFÁTICOS

Lic. María José Castro

Profesor: Dr. Ricardo Maronna

Marzo 2021

1. INTRODUCCIÓN

Introducción al problema

Los alcoholes son sustancias tóxicas y, por lo tanto, constituyen contaminantes ambientales peligrosos. El primer paso en el mecanismo contaminante de los alcoholes es su solubilidad en agua (solubilidad acuosa). Esta es una propiedad física que ha sido ampliamente estudiada. Como propiedad que incluye el agua como solvente, es importante en un conjunto diverso y en un gran número de situaciones, incluidas las aplicaciones químicas, bioquímicas, farmacéuticas, ambientales e industriales.

El desarrollo de una relación o estudio QSPR (*quantitative structure property relationship*) es útil para comprender la solubilidad acuosa y puede proporcionar un método suficientemente simple para estimar el valor de esta propiedad directamente desde la estructura química sin ser necesario recurrir a la medición experimental.

El aspecto clave en un estudio QSPR es desarrollar un modelo matemático que relacione un conjunto de descriptores con una propiedad. Este conjunto está compuesto por un amplio número de parámetros característicos que abarcan propiedades físicas y químicas significativas, permitiendo muchas y variadas posibilidades de elección.

Objetivo del trabajo

El objetivo de este trabajo es diseñar un modelo matemático que permita predecir la solubilidad en agua de alcoholes alifáticos en función de una selección de descriptores fisicoquímicos y comparar con los resultados publicados por Romanelli *et al*.(2001).

Materiales y métodos

Para predecir la solubilidad en agua de alcoholes alifáticos se consideraron los siguientes descriptores:

- Volumen molecular limitado a la superficie accesible al solvente (**SAG**),
- Volumen (**V**),
- Polarizabilidad[1] (**P**),
- Logaritmo del coeficiente de partición octanol-agua[2] **(logPC)**
- Refractividad molar[3] (**RM**), y
- Masa molar (**Mass**).

Los datos del problema se encuentran en la Tabla B1 del Apéndice B y corresponden a los publicados por Romanelli *et al*. (2001). La base de datos contiene información de 7

[1] Facilidad con que puede distorsionarse la nube electrónica de un átomo o molécula; cuanto mayor es la masa molar (**Mass**) más polarizable es la molécula.

[2] Caracteriza la hidrofobicidad de una molécula: a mayor PC (coeficiente de partición), mayor hidrofobicidad y menor solubilidad en agua; depende de la temperatura, del volumen (**V)**, de la masa molar (**Mass)** y de la polarizabilidad (**P**).

[3] Variable relacionada con la temperatura, el índice de refracción, la presión y la masa molar. A mayor masa molar, mayor refractividad.

variables continuas para 44 alcoholes alifáticos: el logaritmo de la solubilidad en agua (**logSolubility**) y los valores correspondientes a los 6 descriptores elegidos.

Se realizó un análisis de las variables involucradas en el problema. Se propuso un modelo aditivo con todos los descriptores fisicoquímicos como variables predictoras. El ajuste del modelo se realizó utilizando dos métodos: mínimos cuadrados ordinarios y métodos robustos. Se llevó a cabo un diagnóstico del ajuste de mínimos cuadrados para analizar la presencia de posibles outliers y violaciones de los supuestos necesarios para hacer inferencia a partir de los resultados de mínimos cuadrados. Se realizó una comparación entre los ajustes propuestos. También se realizó un breve análisis de los modelos publicados por Romanelli *et al*. (2001).

En el Apéndice A se resume la teoría utilizada en el trabajo, y en el Apéndice B se encuentran los datos utilizados. Se fijó el nivel de significación α (error de tipo I: p-valor) en $0{,}05$. En el caso de las pruebas utilizadas para el estudio de normalidad y homocedasticidad se fijó en $0{,}15$. El software utilizado fue R Core Team (2020)

2. MODELO MATEMÁTICO PARA ESTIMAR LA SOLUBILIDAD ACUOSA DE ALCOHOLES ALIFÁTICOS

Análisis descriptivo de las variables del problema

La **Figura 1** muestra el resumen de las estadísticas descriptivas y los gráficos de boxplot correspondientes a los descriptores fisicoquímicos (variables predictoras).

Medidas descriptivas	**SAG**	**V**		**logPC**	**P**	**RM**	**Mass**
Mínimo	247,60	344,90		0,94	8,75	21,95	74,12
1er cuarto	290,40	429,70		1,68	12,42	31,07	102,18
Mediana	312,30	472,00		2,05	14,26	35,63	116,20
Media	337,80	509,60		2,25	14,88	37,33	120,99
3er cuarto	371,30	565,10		2,87	17,93	44,79	144,26
Máximo	587,00	938,50		5,30	28,94	72,75	228,42

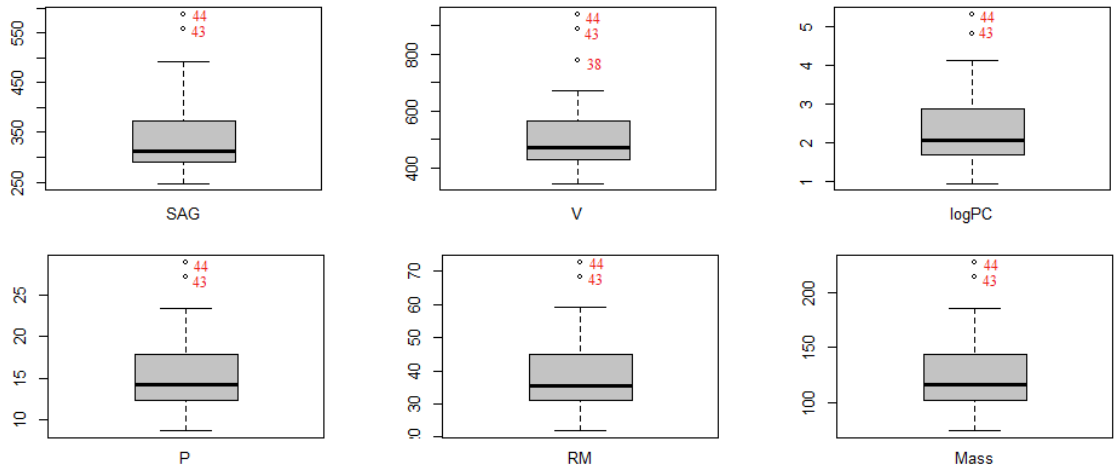

Figura 1: Resumen de estadísticos descriptivos y boxplot de variables predictoras

Respecto al tipo de distribución de estas variables se observa una simetría a la derecha en todos los casos, identificándose (en todas las variables) las observaciones 43 y 44 con los valores más extremos. Las variables predictoras (de a pares) se encuentran altamente correlacionadas (**Tabla 1**).

	SAG	V	logPC	P	RM	Mass
SAG	1	0,9970	0,9740	0,9784	0,9800	0,9784
V		1	0,9863	0,9911	0,9921	0,9911
logPC			1	0,9934	0,9924	0,9934
P				1	0,9998	1
RM					1	0,9998
Mass						1

Tabla 1: Coeficiente de correlación (r) de las variables predictoras

Los datos correspondientes a la variable respuesta **logSolubility** se distribuyen en forma asimétrica, cola pesada a la izquierda. Nuevamente las observaciones 43 y 44 tienen los valores más extremos (**Figura 2**).

	logSolubility
Mínimo	-14,6140
1er cuarto	-5,6157
Mediana	-2,7526
Media	-3,7108
3er cuarto	-1,5438
Máximo	0,3386

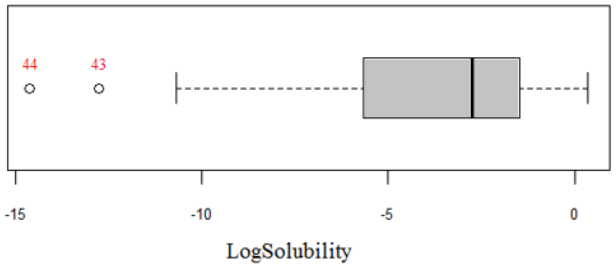

Figura 2: Estadísticas descriptivas y boxplot variable respuesta logSolubility

La variable respuesta **logSolubility** se encuentra altamente correlacionada con cada una de las variables predictoras (**Gráfico 3**).

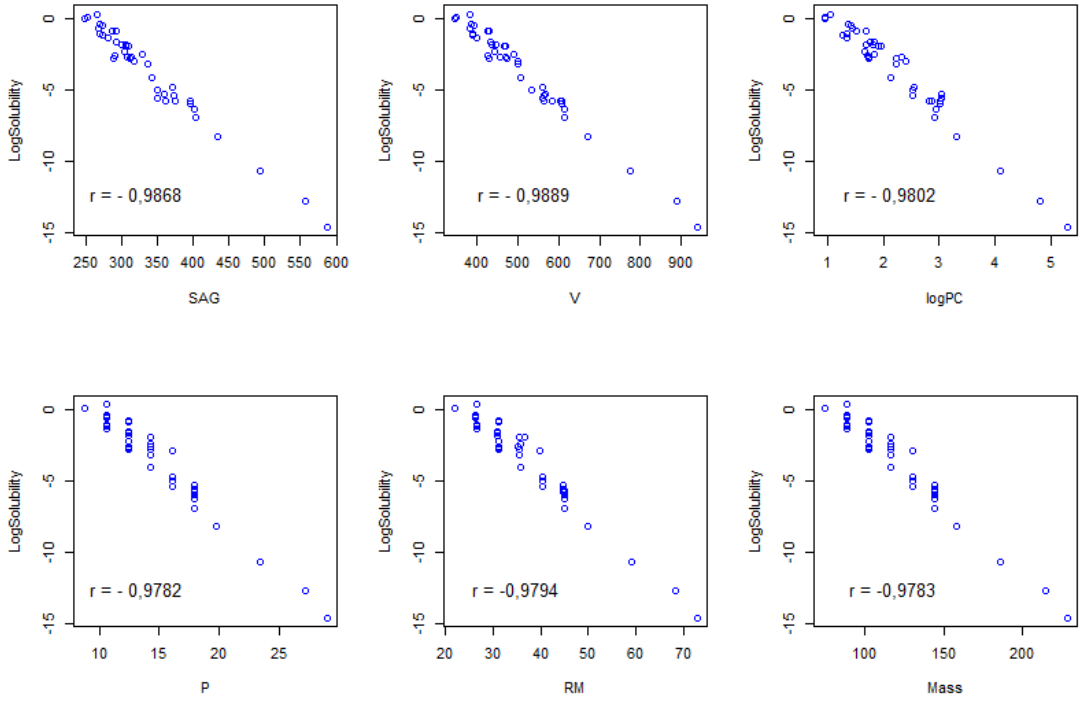

Gráfico 3: logSolubility vs variables predictoras

Modelo matemático para el problema

Se propuso un modelo aditivo para predecir el logaritmo de la solubilidad acuosa que incluye todas las variables predictoras.

$$\text{logSolubility} = \beta_0 + \beta_1 \cdot \text{SAG} + \beta_2 \cdot \text{V} + \beta_3 \cdot \text{logPC} + \beta_4 \cdot \text{P} + \beta_5 \cdot \text{RM} + \beta_6 \cdot \text{Mass} + u$$

u: término de error (aleatorio)

β_i: parámetros del modelo, $\quad i = 0, \ldots, 6$

Para la estimación de los parámetros del modelo se utilizaron dos metodologías: mínimos cuadrados clásicos y métodos robustos, realizándose un diagnóstico del modelo en el caso de mínimos cuadrados.

3.2.1 Estimación de los coeficientes del modelo completo por mínimos cuadrados

Los resultados obtenidos por este método (*ajusteLS*) se encuentran en la **Salida 1**.

Variables (coeficientes)	**Estimador**	**(E.S)**
Intercept (β_0)	35,868	(25,652)
SAG (β_1)	-0,005	(0,046)
V (β_2)	-0,018	(0,041)
logPC (β_3)	-2,396*	(0,746)
P (β_4)	32,989	(27,218)
RM (β_5)	-0,673	(0,418)
Mass (β_6)	-4,045	(3,559)
Observaciones	44	
R^2	0,983	
R^2 ajustado	0,980	
Residual Std. Error	0,459 (df = 37)	
EstadísticoF	361.327* (df = 6; 37)	
* p < 0,05		
( ) : error estándar del estimador		

Salida 1: Estimación de los coeficientes por mínimos cuadrados (*ajusteLS*)

Globalmente, el ajuste resulta significativo ($p < 0,05$); solo la variable **logPC** resulta significativa ($p < 0,05$) y el coeficiente estimado correspondiente a la variable **P** es el único con signo positivo (era de esperarse signo negativo **Gráfico 3**).

El gráfico de los residuos del modelo versus cada una de las variables predictoras no muestra evidencia para suponer que no existe relación lineal entre la variable respuesta y las variables predictoras (**Gráfico 4**).

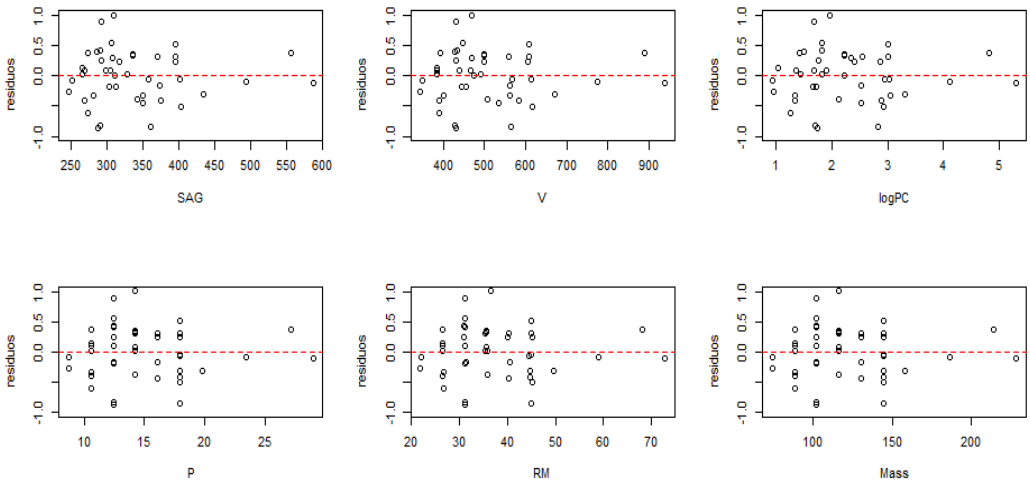

Gráfico 4: Gráfico de los residuos vs variables predictoras (*ajusteLS)*

Se realizó el diagnóstico del modelo ajustado por mínimos cuadrados analizando el supuesto de normalidad y homocedasticidad de los errores ($u \sim N(0; \sigma^2 I)$), como también la posible presencia de datos atípicos/influyentes y multicolinealidad.

El QQ-Plot de los residuos estudentizados (**Gráfico 5**) muestra a la observación 39 con residuo alto ($\gg 2{,}5$). El test de Kolmogorov-Smirnov para el análisis del supuesto de normalidad de los residuos es no significativo ($p = 0{,}16$).

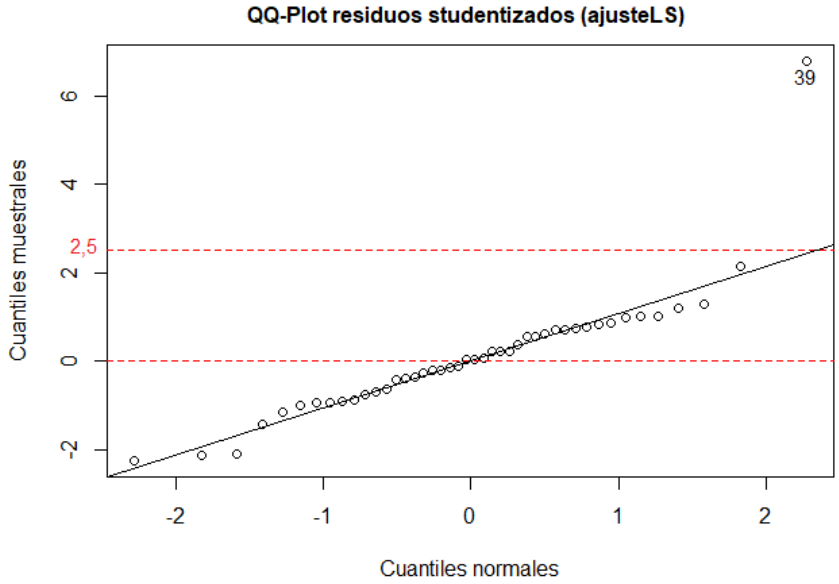

Gráfico 5: QQ-Plot de los residuos estudentizados (*ajusteLS)*

El **Gráfico 6** correspondiente al valor absoluto de los residuos estudentizados vs predichos muestra una posible estructura: a mayor predicho, mayor valor absoluto de residuo (en el gráfico se excluye la observación 39 por su alto residuo.). El test de Breusch-Pagan resultó significativo ($p = 0{,}001$) rechazándose la hipótesis de homocedasticidad de los residuos.

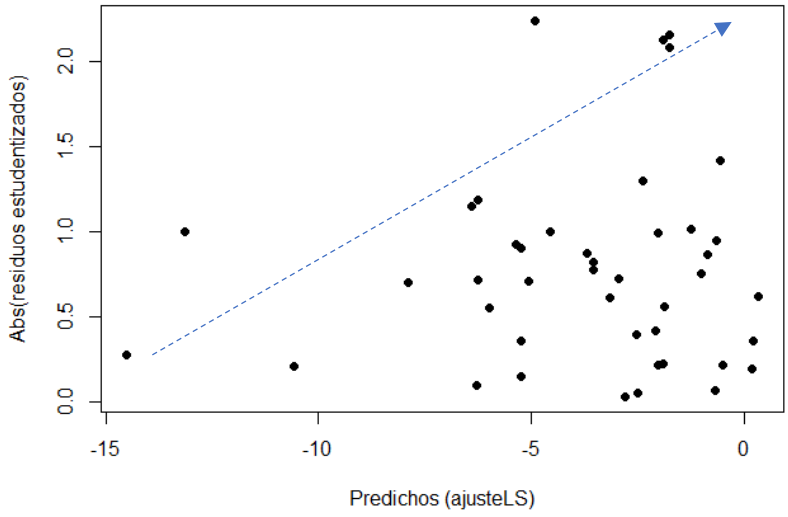

Gráfico 6: Abs(residuos estudentizados) vs predichos

Respecto a la presencia de valores atípicos y/o influyentes, se observa en el **Gráfico 7** a las observaciones 39 y 41 con alto leverage (mayor a 0,40[4]).

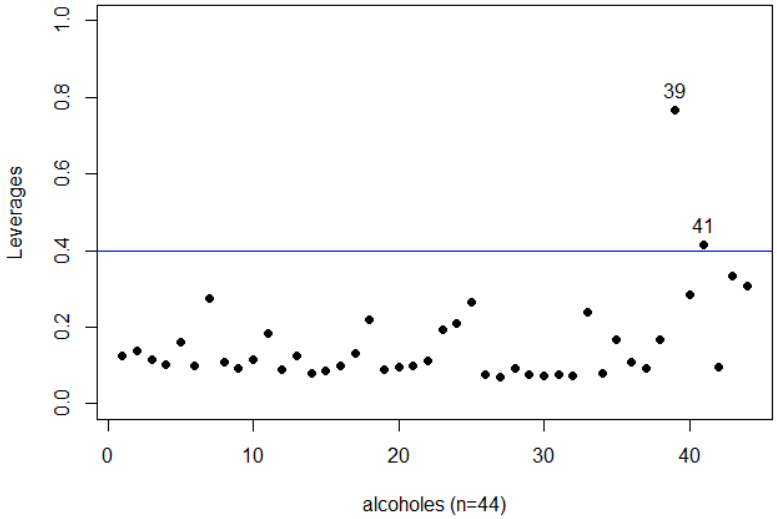

Gráfico 7: Leverage de las observaciones (*ajusteLS*)

El análisis de las distancias de Cook (**Gráfico 8**) confirma a la observación 39 como influyente.

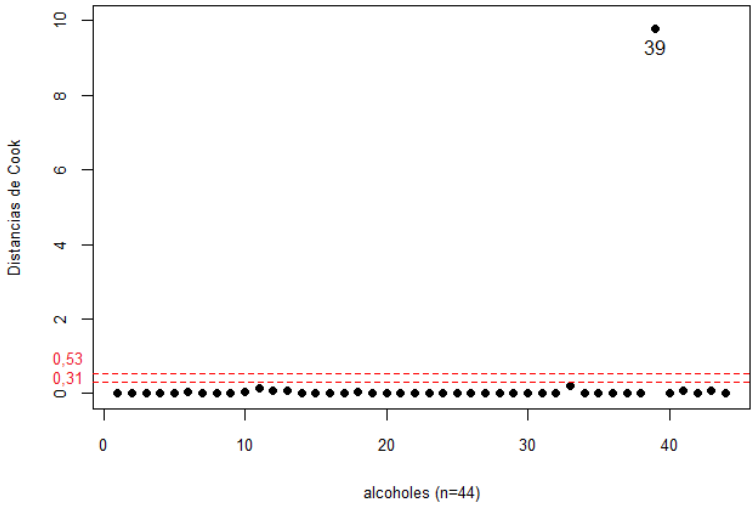

Gráfico 8: Distancias de Cook (*ajusteLS*)

Se analizó la influencia que tienen las observaciones 39, 41 y 43 en la estimación de los coeficientes del modelo por mínimos cuadrados (se excluyó una observación a la vez y se estimó los coeficientes del modelo en cada caso). Se observa una gran diferencia respecto a los coeficientes correspondientes al *ajusteLS* cuando se excluye la observación 39; no así cuando se excluyen las observaciones 41 y 43 (**Tabla 2**).

[4] Cota leverage: $0{,}40 = 2{,}5 \cdot \frac{p}{n}$, con $p = 7$ y $n = 44$

	Intercept	SAG	V	logPC	P	RM	Mass
ajusteLS	35.8677	-0.0047	-0.0177	-2.3960	32.9889	-0.6726	-4.0453
ajusteLS [-obs 39]	-4.2939	-0.0067	0.0088	-3.7497	1.5367	-3.8361	1.0499
ajusteLS [-obs 41]	38.0495	-0.0291	0.0017	-2.2310	33.8800	-0.5789	-4.2188
ajusteLS [-obs 43]	40.4597	0.0082	-0.0302	-2.2304	37.8975	-0.6531	-4.6800

Tabla 2: Comparación de los coeficientes (*ajusteLS*)

Los factores de inflación de la varianza, VIF, muestran (**Tabla 3**) un grave problema de multicolinealidad, cada una de las variables predictoras tiene un VIF mucho mayor a 10.

VIF	SAG	V	logPC	P	RM	Mass
	2.376,8	5.724,2	100,80	2.821.432,6	4.208,0	2.819.048,8

Tabla 3: valores VIF (*ajusteLS*)

Conclusión del diagnóstico del modelo (*ajusteLS*):

- no hay evidencia para rechazar el supuesto de normalidad de los errores,
- los errores no son homocedásticos,
- presencia de outliers: observación 39,
- problema de multicolinealidad.

Estimación de los coeficientes del modelo completo por métodos robustos

Se realizó un primer ajuste robusto (*ajusteMM*) utilizando un MM-estimador y en segundo lugar un DCML-estimador. En ambos casos se consideró una eficiencia normal del 85%, con punto de corte 0.5 y familia bicuadrada. La **Salida 2** muestra los resultados obtenidos.

Variable (coeficiente)	*ajusteMM*	*ajusteDCML*
Intercept (β_0)	-1,7576 (14,988)	4,9400 (13,3412)
SAG (β_1)	-0,0412 (0,027)	-0,0347 (0,024)
V (β_2)	-0,0420 (0,024)	-0,0314 (0,021)
logPC (β_3)	-3,8766* (0,746)	-3,6130* (0,384)
P (β_4)	3,9600 (15,574)	9,1274 (13,8621)
RM (β_5)	-4,1138* (0,452)	-3,5012* (0,4023)
Mass (β_6)	0,7754 (2,065)	-0,0827 (1,8377)
Observaciones	44	44
Factor de escala robusto	0,2546	0,2680
* p < 0,05	( ) : error estándar de la estimación	

Salida 2: Estimaciones de los coeficientes por métodos robustos

El *ajusteDCML* proporciona mejores resultados que el *ajusteMM*: los errores de las estimaciones son más pequeños y la suma de cuadrados de los residuos es menor (19,7153 y 25,4433 respectivamente).

Ajuste robusto *ajusteDCML*

El análisis de los residuos del *ajusteDCML* (**Gráfico 9**) muestra que, con excepción de las observaciones 12 y 39 los residuos, en módulo, toman valores menores o iguales a 0,6070 (2,5 veces el factor de escala robusto *s* = 0,2680). La observación 36 se encuentra en el límite (0,5875).

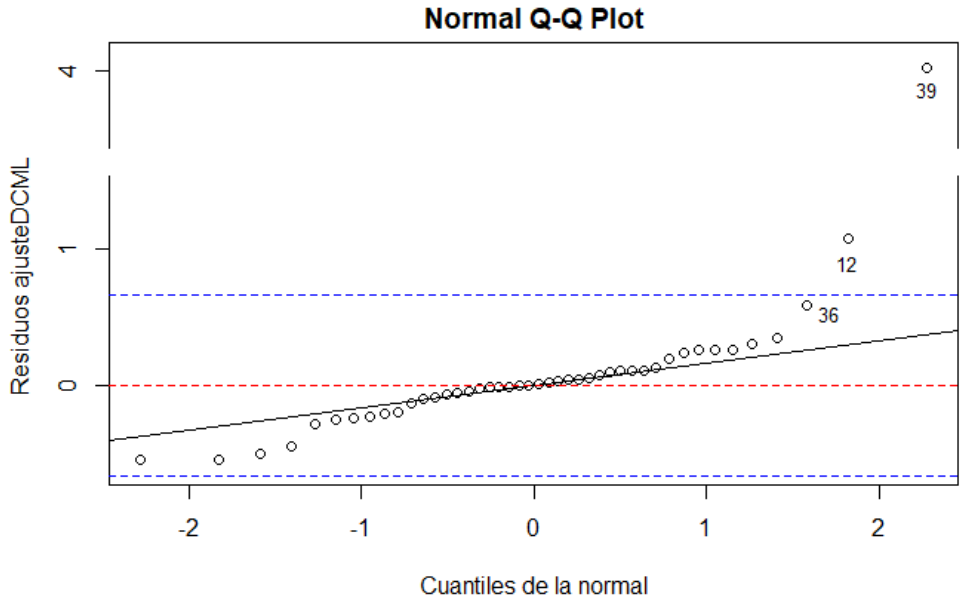

Gráfico 9: QQ-Plot residuos *ajusteDCML*

La **Tabla 4** muestra los pesos que el ajuste robusto asigna a cada una de las observaciones: las observaciones 12 y 39 reciben peso cero; otras observaciones que reciben poco peso son las número 36, 2 y 13.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
0.9161	0.3434	0.9844	0.8557	0.8934	0.9587	0.9729	0.9956	0.9965	0.7642	0.6716
12	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
0.0000	0.4914	0.9716	0.9855	0.9961	0.9951	0.8878	0.9986	0.8397	0.8486	0.7664
23	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**
0.9987	0.9978	0.9897	0.9997	0.9922	0.9818	0.8602	0.8152	0.9992	0.8233	0.9984
34	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**
0.3082	0.9692	0.1692	0.9256	0.9946	0.0000	0.9861	0.7612	0.9812	0.9944	0.9949

Tabla 4: Pesos para cada observación (*ajusteDCML*)

Ajuste de mínimos cuadrados vs ajuste robusto con un estimador *DCML*

Para el conjunto de datos, los resultados que se obtienen con el ajuste de mínimos cuadrados difieren de los obtenidos con el ajuste robusto: la estimación de los coeficientes del modelo cambia, al igual que la significatividad de las variables; y los errores de las estimaciones del método robusto son menores (**Salida 3**).

Variable (coeficiente)	***ajusteLS***	***ajusteDCML***
Intercept (β_0)	35,8677 (25,652)	4,9400 (13,341)
SAG (β_1)	- 0,0047 (0,046)	- 0,0347 (0,024)
V (β_2)	- 0,0177 (0,041)	- 0,0314 (0,021)
logPC (β_3)	- 2,3960* (0,746)	-3,6130* (0,384)
P (β_4)	32,9889 (27,218)	9,1274 (13,8621)
RM (β_5)	- 0,6726 (0,418)	- 3,5012* (0,402)
Mass (β_6)	- 4,0453 (3,559)	- 0,0827 (1,838)
* p < 0,05	( ) : error estándar de la estimación	

Salida 3: Estimadores de los coeficientes *ajusteLS* vs *ajusteDCML*

En el **Gráfico 10** vemos que los residuos del ajuste robusto (*ajusteDCML*) son más pequeños que los residuos del ajuste de mínimos cuadrados (*ajusteLS*) salvo en las observaciones con peso cero en el ajuste robusto (12 y 39). El *ajusteDCML* da un mejor ajuste a la mayoría de las observaciones.

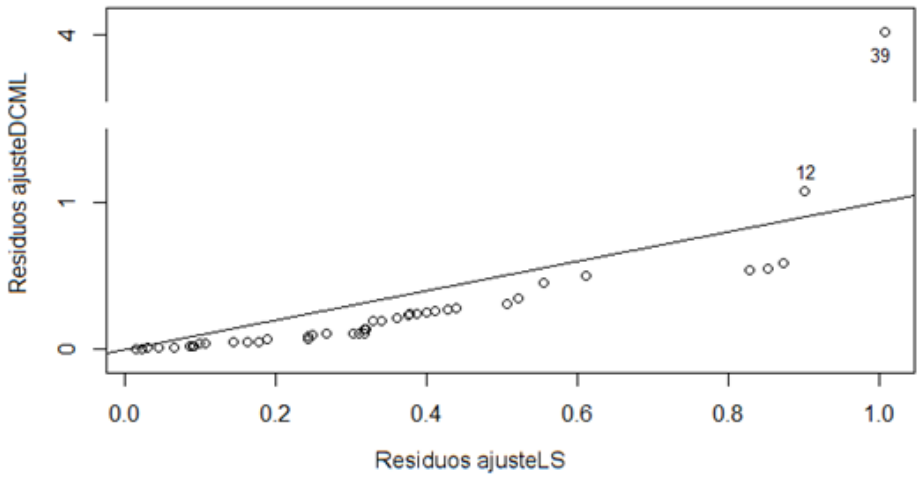

Gráfico 10: Valor absoluto de los residuos *ajusteLS* vs residuos *ajusteDCML* (ordenados de menor a mayor)

Conclusiones

El ajuste del modelo por mínimos cuadrados sólo detecta como outlier la observación 39 mientras que el ajuste robusto detecta un segundo outlier: la observación 12. Al excluir las observaciones 12 y 39 se observa que la suma del cuadrado de los residuos (**Tabla 5**) cae bruscamente con el ajuste robusto, no así en el ajuste por mínimos cuadrados.

	ajusteLS	***ajusteDCML***
Alcohol (n = 44)	7,7823	19,7153
Alcohol [-c(12,39)] (n=42)	5,9547	2,3690
Variación de la suma del cuadrado de los residuos al excluir las observaciones 12 y 39 (GAP)	- 23,5%	- 88,0%

Tabla 5: Suma del cuadrado de los residuos

Si bien la suma del cuadrado de los residuos del ajuste robusto es mayor que la correspondiente al ajuste de mínimos cuadrados, cuando se excluyen de la suma las observaciones 12 y 39 se produce una disminución en dicho valor del 88,0% en el caso del ajuste robusto y solo un 23,5% en el ajuste de mínimos cuadrados. Claramente el ajuste por mínimos cuadrados se ve afectado por los outliers.

Se reprodujeron los modelos propuestos por Romanelli *et al.* (2001). En dicho trabajo no se mencionó la presencia de outliers, ni el problema de multicolinealidad; en el caso de los modelos 8 y 9 no se tuvo en cuenta la falta de normalidad de los errores.

Mod1: logSolubility ~SAG
Mod2: logSolubility ~V
Mod3: logSolubility ~$V+V^2+V^3$
Mod4: logSolubility ~$SAG+SAG^2$
Mod5: logSolubility ~V+logPC
Mod6: logSolubility ~$SAG+P+SAG^2+P^2$
Mod7: logSolubility ~$SAG+RM+SAG^2+RM^2+SAG^3+RM^3$
Mod8: logSolubility ~P+RM+Mass
Mod9: logSolubility ~SAG+logPC+RM

En la **Tabla 10** resume la suma de los cuadrados de los residuos para cada uno de los modelos donde se evidencia el mismo efecto que en el *ajusteLS*: la exclusión de las observaciones 12 y 39 produce un GAP que oscila, en módulo, entre 1,7% y 33,2% siendo mucho menor al GAP que se logra con el ajuste robusto.

	Mod1	**Mod2**	**Mod3**	**Mod4**	**Mod5**	**Mod6**	**Mod7**	**Mod8**	**Mod9**
Alcohol (n = 44)	12,1338	10,2676	9,8798	11,8962	9,8547	8,9642	8,7152	17,2939	8,7243
Alcohol [-c(12,39)] (n=42)	11,0863	8,9155	8,7267	10,8693	8,5015	8,1081	7,8645	11,5471	7,7837
GAP	**-8,6%**	**-13,2%**	**-1,7%**	**- 8,6%**	**-13,7%**	**-9,5%**	**-9 ,8%**	**-33,2%**	**-10,8%**

Tabla 6: Suma de los cuadrados de los residuos – Modelos trabajo Romanelli *et al.* (2001)

Se propone como modelo final para estimar el logaritmo de la solubilidad de alcoholes alifáticos el que se obtiene al utilizar la totalidad de los datos de la muestra, y la estimación de los coeficientes del modelo con el *ajusteDCML*

$$\begin{aligned}\text{logSolubility} = {} & 4{,}94001 - 0{,}03474 \cdot \text{SAG} + 0{,}03136 \cdot \text{V} - 3{,}61303 \cdot \text{logPC} \\ & + 9{,}12737 \cdot \text{P} - 3{,}50123 \cdot \text{RM} - 0{,}08273 \cdot \text{Mass}\end{aligned}$$

Apéndice A

Modelo clásico de regresión lineal múltiple. Algunas definiciones.

Cuando hablamos de "diseñar un modelo" nos referimos a hallar una expresión o función matemática que permita describir el comportamiento de la variable de interés (variable respuesta) en función de otras variables (predictoras o regresoras).
El modelo de regresión lineal múltiple es uno de los más utilizados entre los modelos estadísticos. En este modelo, la función que relaciona la variable respuesta y con las variables predictoras $(x_1, x_2, \cdots, x_p)$ es lineal, es decir, es de la forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \text{u} \quad (2.1)$$

donde $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ son constantes desconocidas denominadas parámetros o coeficientes de regresión, y u es el término de error no observable (variable aleatoria con cierta distribución).

Para estimar los coeficientes del modelo se necesitan n observaciones $(n > p)$. Cuando las variables predictoras son aleatorias, es decir, son observadas al mismo tiempo que la variable respuesta, nuestro conjunto de observaciones está formado por vectores aleatorios de la forma $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i) = (\mathbf{x}_i', y_i)$ con $1 \leq i \leq n$ que satisfacen la relación lineal

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \ + u_i \quad 1 \leq \text{i} \leq \text{n} \quad (2.2)$$

Si definimos la matriz de diseño **X**, y los vectores **y**, **β**, **u**

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_1' \\ \vdots & \vdots \\ 1 & \mathbf{x}_i' \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix},$$
$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

el modelo se puede expresar como

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u} \quad (2.3).$$

Sea $\widehat{\boldsymbol{\beta}}$ un vector conocido de coeficientes. Los valores $\hat{y}_i$ (predicho) y r_i (residuo) para $1 \leq i \leq n$, asociados a dicho vector $\widehat{\boldsymbol{\beta}}$, se definen como

$$\hat{y}_i = \hat{y}_i(\widehat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \; r_i = r_i(\widehat{\boldsymbol{\beta}}) = y_i - \hat{y}_i \qquad (2.4).$$

En forma matricial tenemos

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \qquad \hat{\boldsymbol{r}} = \mathbf{y} - \hat{\mathbf{y}}$$

con $\hat{\mathbf{y}}^t = (\hat{y}_1, \ldots, \hat{y}_m)$, $\hat{\mathbf{r}}^t = (\hat{r}_1, \ldots, \hat{r}_m)$.

Sea

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

la matriz de $\mathbb{R}^{n\times n}$ y sea $h_i = [\mathbf{H}]_{ii}$ el i-ésimo elemento de la diagonal $(1 \leq i \leq n)$ llamado también *leverage*. El i-ésimo residuo definido en (2.4) se puede expresar en función de la matriz **H** como

$$r_i = (1 - H)\, u_i.$$

Si además suponemos que $\mathrm{E}(u_i) = 0$ y $\mathrm{Var}(u_i) = \sigma^2$ se tiene que

$$Var(r_i) = \sigma^2(1 - h_i).$$

A partir del residuo clásico pueden definirse otros residuos: residuo de *cross validation* y el residuo *estudentizado*.

El residuo de "*cross validation*" r_{-i} se define como

$$r_{-i} = r_{-i}\left(\widehat{\boldsymbol{\beta}}_{-\boldsymbol{i}}\right) = y_i - \hat{y}_{-i},$$

donde $\widehat{\boldsymbol{\beta}}_{-\boldsymbol{i}}$ es el estimador de mínimos cuadrados calculado sin usar la i -ésima observación y el valor $\hat{y}_{-i}$ es el predicho por el modelo para dicha observación. Se puede mostrar que

$$r_{-i} = \frac{r_i}{1 - h_i}$$

siendo la versión normalizada del *residuo cross validation* es

$$t_i = \frac{r_i}{s\sqrt{1 - h_i}} \quad ; \quad s^2 = \hat{\sigma}^2 = \frac{1}{n - p^*}\sum_{i=1}^{n} r_i^2,$$

donde $p^* = \mathrm{rango}(\mathbf{X})$ y la matriz **X** puede ser o no de rango completo.

El *residuo studentizado* t_{-i} se define como

$$t_{-i} = \frac{r_i}{s_{-i}\sqrt{1 - h_i}} \quad ; \quad s_{-i}^2 = \frac{1}{(n - p^* - 1)}\sum_{i=1}^{n} r_{-i}^2.$$

Para el caso en que matriz **X** es aleatoria, un estimador del error cuadrático medio EMC para el estimador $\widehat{\boldsymbol{\beta}}$ se define como

$$\mathrm{E}\widehat{M}\mathrm{C}\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i^2}{(1 - h_i)^2}.$$

Estimación de los coeficientes del modelo: mínimos cuadrados.

El método clásico para estimar el vector **β** es el de mínimos cuadrados (LS): este vector, que llamamos, $\widehat{\boldsymbol{\beta}}_{\mathbf{LS}}$ es el que minimiza la suma

$$\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}(\boldsymbol{r_i}) \text{ con } \rho(t) = t^2 \qquad (2.5)$$

De los errores u_i supondremos que son variables aleatorias independientes igualmente distribuidas, con $\mathrm{E}(u_i) = 0, \mathrm{Var}(u_i) = \sigma^2$ y también independientes de las variables $\mathbf{x}_i' = (x_{i1}, x_{i2}, \dots, x_{ip})$ con $1 \leq i \leq n$ que siguen cierta distribución.

Si la distribución del vector **x** no está centrada en ningún hiperplano, es decir,

$$P(a^t\mathbf{x} = 0) < 1 \quad \forall\, a \neq 0$$

se tiene que la probabilidad de que la matriz **X** tenga rango completo tiende a 1 cuando $n \to \infty$, y se cumple en particular si **x** tiene densidad. En este caso, el estimador de mínimos cuadrados está bien definido y se expresa como

$$\widehat{\boldsymbol{\beta}}_{\mathbf{LS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Las expresiones para la esperanza y varianza de $\widehat{\boldsymbol{\beta}}_{\mathbf{LS}}$ son válidas condicionalmente, esto es

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\mathbf{LS}} | \mathbf{X} = \mathbf{x}) = \boldsymbol{\beta}, \qquad \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathbf{LS}} | \mathbf{X} = \mathbf{x}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Si, además, los u_i se encuentran distribuidos normalmente se tiene que

$$\widehat{\boldsymbol{\beta}}_{\mathbf{LS}} \sim \mathrm{N_p}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Problemas con el estimador de mínimos cuadrados: diagnóstico del modelo.

Si los supuestos sobre los errores se satisfacen y no existe problema de colinealidad (la matriz de diseño tiene rango completo) el estimador de mínimos cuadrados tiene propiedades estadísticas deseables: es insesgado, de mínima varianza entre todos los insesgados y coincide con el estimador de máxima verosimilitud (heredando todas las propiedades de este último).
Cuando la matriz **X** no es de rango completo, estamos frente a un problema de colinealidad. En este caso el sistema (2.5) tiene infinitas soluciones, pero los valores predichos para cada una de las observaciones son los mismos y, por lo tanto, también los residuos.

Si algunos de los supuestos realizados para estimar los coeficientes no son válidos los coeficientes estimados por el método de mínimos cuadrados pueden ser engañosos. En el caso de que la distribución de los errores es de cola pesada, el estimador de mínimos cuadrados ya no tiene las propiedades deseables. Este tipo de distribuciones tienden a generar valores atípicos, los cuales pueden tener un efecto inadecuado en las estimaciones. Observaciones atípicas, tanto en la dirección de la variable repuesta como de las predictoras, pueden tener un fuerte efecto adverso en la estimación de mínimos cuadrados, pudiendo pasar desapercibidos.

La validez de las estimaciones y la inferencia a partir del modelo de regresión múltiple dependen de varios supuestos los cuales deben verificarse antes de poner en práctica el modelo estimado. Éstos incluyen supuestos referidos al término del error $\boldsymbol{u}$, validez del modelo lineal propuesto, no colinealidad y la existencia de observaciones atípicas (en algunas ocasiones, unas pocas observaciones no se ajustan al modelo y pueden cambiar su elección y ajuste). Dado que los errores no son observables, los residuos del modelo son de gran utilidad para el análisis de los supuestos planteados sobre los errores.

Supuestos y herramientas para su validación (ajuste de mínimos cuadrados).

Para poder hacer inferencia a partir de las estimaciones obtenidas con el estimador de mínimos cuadrados, debemos verificar que la $\mathbf{Var}(\boldsymbol{u_i}) = \boldsymbol{\sigma^2}$, que se distribuyan normalmente, que no haya colinealidad, que el modelo lineal sea válido, y analizar la presencia de valores atípicos.

i) Varianza constante

La varianza de los residuos debe de ser constante en todo el rango de observaciones. Si la varianza es constante, el gráfico de los *valores absolutos de los residuos* t_i *vs los valores predichos* por el modelo muestra una disposición aleatoria manteniendo una misma dispersión y sin ningún patrón específico (bajo el supuesto de homocedasticidad, los t_i tiene la misma varianza a diferencia de los r_i). También se puede recurrir a contrastes de homocedasticidad como el test de *Breusch-Pagan.*

ii) Distribución normal

Gráficamente, se puede realizar el QQ-Plot de la versión estandarizada de los *residuos cross validation* y contrastes como el *test de Kolmogorov* Smirnov.

iii) No colinealidad

En presencia de colinealidad, no se puede identificar en forma precisa el efecto individual que tiene cada una de las variables predictoras sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes estimados $\widehat{\boldsymbol{\beta}}$. Esto hace que los intervalos de confianza para los coeficientes sean demasiados amplios y se pierda potencia en los contrastes de hipótesis. Se tenderá a no rechazar la hipótesis de significación estadística de los coeficientes del modelo sin que esto implique que el modelo sea globalmente no significativo. Es decir, se tendría un valor de R^2 alto con pocos o ninguno de los coeficientes del modelo significativos.

Hay varios caminos que nos pueden indicar la posible existencia de colinealidad:

- Si el coeficiente de determinación R^2 es alto, pero ninguno de los predictores resulta significativo.
- Calcular la matriz de correlación en la que se estudia la relación lineal entre cada par de predictores (sólo evalúa correlación de a pares).
- Generar un modelo de regresión lineal simple entre cada uno de los predictores frente al resto: si en alguno de los modelos el coeficiente de determinación R^2 es alto, estaría señalando una posible colinealidad.
- *Factor de Inflación de la Varianza (VIF)*. Se calcula para cada una de las variables predictoras según la siguiente fórmula:

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}$$

donde los R_j^2 se obtiene de la regresión de la variable x_j sobre las otras variables predictoras. Según el valor tenemos: $VIF = 1$: ausencia total de colinealidad; $1 < VIF < 5$: la regresión puede verse afectada por cierta colinealidad; $5 < VIF < 10$: causa de preocupación; y $VIF > 10$: problema grave de colinealidad.

iv) Relación lineal entre los predictores numéricos y la variable respuesta

Cada predictor numérico tiene que estar linealmente relacionado con la variable respuesta mientras los demás predictores se mantienen constantes, de lo contrario no se puede introducir en el modelo. La forma más recomendable de comprobarlo es representando los residuos del modelo frente a cada uno de los predictores. Si la relación es lineal, los residuos se distribuyen de forma aleatoria en torno al cero (este análisis es sólo aproximado).

v) Observaciones atípicas (outliers), influyentes y con alto leverage

Los outliers u observaciones atípicas son aquellas que no se ajustan en forma adecuada al modelo. En estos casos, el valor real de la variable respuesta se aleja del valor predicho por el modelo dando lugar a un residuo grande.

Una observación influyente es aquella cuya exclusión afecta al modelo. La eliminación de este tipo de observaciones debe analizarse con detalle y si el objetivo del modelo es predecir futuros valores de la variable respuesta, un modelo sin estas observaciones puede ser más útil para predecir con precisión la mayoría de los casos. De todas maneras, es muy importante prestar atención a estos valores ya que, de no ser errores de medida, pueden ser los casos más interesantes.

Las observaciones con alto *leverage* son aquellas con valores extremos para alguna de las variables predictoras y son potenciales puntos influyentes. Observaciones donde $h_i > 2{,}5 \cdot p/n$ (siendo p la cantidad de predictores más el término independiente y n la cantidad de observaciones) deben tenerse en consideración por su posible influencia.

La detección de la presencia de outliers también puede hacerse utilizando los residuos estudentizados t_{-i}. En un gráfico QQ-Plot, los puntos que se alejan de la recta $y = x$ son posibles outliers. También se pueden considerar sospechosas (sin hacer ningún tipo de gráfico) aquellas observaciones con $|t_{-i}| > 2{,}5$

El hecho de que un valor sea atípico o con alto grado de *leverage* no implica que sea influyente en el conjunto del modelo. Sin embargo, si un valor es influyente, suele ser o atípico o de alto *leverage*. Otra forma de evaluar la influencia de las observaciones es utilizando la *distancia de Cook*. Es una medida que combina, en un único valor, la magnitud del residuo y el grado de *leverage*. Observaciones con valores de *distancia de Cook* mayor a 1 suelen considerarse como influyentes.

La regresión robusta es un método importante para detectar la presencia de valores atípicos y proporcionar resultados resistentes, ya que en muchos casos la presencia de estos datos puede estar enmascarada.

Estimación de los coeficientes del modelo: métodos robustos.

Cuando los supuestos del modelo no se cumplen por la naturaleza de los datos, las técnicas de regresión robusta proporcionan una alternativa al método de mínimos cuadrados. El objetivo principal de estas técnicas es obtener estimadores de los coeficientes que puedan proporcionar resultados resistentes a la presencia de valores atípicos. Es decir, ajustar un modelo que describa de la mejor forma la mayoría de los datos de la muestra.

La robustez del procedimiento se logra al asignar pesos a los datos dentro del cálculo de la estimación, de modo que los posibles outliers tengan una menor influencia en las estimaciones de los coeficientes.

Es de esperarse que estas técnicas funcionen bien tanto en la presencia como ausencia de valores atípicos; y sean una herramienta más para la detección de posibles outliers en conjunto de datos con mayor complejidad.

Se propusieron muchos estimadores de regresión robustos, y para obtener una apreciación de sus fortalezas y debilidades es necesario de definir algunas propiedades clave: punto de ruptura y eficiencia.

- **Punto de ruptura**
  Se define como el punto o porcentaje mínimo de datos atípicos que puede hacer que el estimador no sea útil. En el caso del estimador de mínimos cuadrados es $1/n$ para una muestra de tamaño n, es decir, una sola observación puede distorsionar el estimador. Esto tiene un impacto potencialmente grave sobre su uso práctico. Si la técnica de estimación robusta tiene un 50% punto de ruptura, entonces el 50% de los datos podrían contener valores atípicos y los coeficientes seguirían siendo utilizables.

- **Eficiencia**
  La eficiencia del estimador robusto se define como el cociente entre el cuadrado medio residual obtenido con el estimador de máxima verosimilitud (el que coincide con el estimador de mínimos cuadrados cuando los supuestos son válidos) y el cuadrado medio residual obtenido con el procedimiento robusto. Es esperable que esta medida de eficiencia se aproxime a 1 cuando $n \to \infty$.
  La mayoría de los resultados teóricos sobre la eficiencia de los estimadores robustos se refieren a su eficiencia asintótica: relación entre las varianzas asintóticas del estimador de máxima verosimilitud y del estimador robusto. Sin embargo, a menos que el tamaño de la muestra sea lo suficientemente grande, la eficiencia (obtenida con la muestra) puede ser menor que la eficiencia asintótica.

Posibles estimadores robustos para el modelo de regresión lineal:

Para mayor detalle sobre la construcción de estos dos estimadores ver Maronna (2019).

- **Estimador MM**: propuesto por Yohai (1987). Es un estimador que combina un alto punto de ruptura (0,5) y alta eficiencia asintótica cuando puede suponerse que los errores tienen distribución normal. Está definido por un procedimiento de tres pasos: paso 1) se busca un estimador robusto consistente, con alto punto de ruptura, pero no necesariamente eficiente; paso 2) se busca un M-estimador de escala de los errores usando el estimador del paso anterior; y paso 3) se calcula un M-estimador de regresión (estimador de los coeficientes del modelo lineal) basado en una función descendente psi. Para el cálculo de este estimador se utilizó la función lmrobdetMM de la librería RobStatTM en R.
- **Estimador DCML**: propuesto por Maronna y Yohai (2014), es un estimador con alta eficiencia y robustez cuando el número de observaciones es pequeño para modelos paramétricos. Dado un estimador robusto (por ejemplo, un estimador MM), el estimador DCML se obtiene maximizando la probabilidad condicionada a que la distancia entre los parámetros sea menor que cierto valor dado.

Para el cálculo de este estimador se utilizó la función lmrobdetDCML de la librería RobStatTM en R.

Apéndice B

Observación	SAG	V	logPC	P	RM	Mass	logSolubility
1	251,94	348,23	0,94	8,75	22,13	74,12	0,0953
2	247,55	344,91	0,96	8,75	21,95	74,12	0,0658
3	281,60	401,41	1,34	10,59	26,74	88,15	-1,3471
4	273,15	392,64	1,43	10,59	26,48	88,15	-0,4861
5	268,75	389,56	1,34	10,59	26,61	88,15	-1,0584
6	273,54	389,93	1,27	10,59	26,68	88,15	-1,1796
7	266,07	383,33	1,04	10,59	26,59	88,15	0,3386
8	269,30	385,05	1,36	10,59	26,42	88,15	-0,4050
9	312,98	455,53	1,73	12,42	31,34	102,18	-2,7181
10	306,26	446,43	1,82	12,42	31,08	102,18	-1,8326
11	290,96	427,07	1,71	12,42	31,16	102,18	-2,5903
12	292,28	430,59	1,68	12,42	31,15	102,18	-0,8510
13	288,65	430,29	1,74	12,42	31,21	102,18	-2,7871
14	303,32	443,40	1,67	12,42	31,28	102,18	-2,2828
15	293,04	432,16	1,76	12,42	31,02	102,18	-1,6399
16	299,38	438,27	1,69	12,42	31,10	102,18	-1,8140
17	291,80	434,04	1,83	12,42	30,95	102,18	-1,6094
18	286,68	427,75	1,51	12,42	31,11	102,18	-0,8301
19	342,35	508,64	2,13	14,26	35,94	116,20	-4,0745
20	335,69	499,42	2,22	14,26	35,68	116,20	-3,1942
21	336,59	500,42	2,22	14,26	35,68	116,20	-3,1966
22	311,71	473,87	2,23	14,26	35,42	116,20	-2,8018
23	307,49	470,06	2,34	14,26	35,35	116,20	-2,6437
24	305,76	467,78	1,91	14,26	35,59	116,20	-1,9379
25	328,11	491,34	1,83	14,26	35,79	116,20	-2,4734
26	373,75	563,02	2,53	16,09	40,54	130,23	-5,4015
27	370,46	559,98	2,54	16,09	40,36	130,23	-4,7560
28	350,77	535,64	2,53	16,09	40,41	130,23	-4,9967
29	403,26	615,97	2,92	17,93	45,14	144,26	-6,9078
30	401,55	613,37	2,94	17,93	44,96	144,26	-6,3200
31	395,82	608,30	3,01	17,93	44,88	144,26	-5,9522
32	396,25	607,83	3,01	17,93	44,88	144,26	-5,7446
33	360,85	564,46	2,83	17,93	44,91	144,26	-5,7699
34	374,83	584,21	2,88	17,93	44,78	144,26	-5,7764
35	358,77	566,85	3,03	17,93	44,62	144,26	-5,2983
36	395,75	604,55	2,86	17,93	45,09	144,26	-5,7446
37	434,80	670,25	3,32	19,76	49,74	158,28	-8,2208
38	493,36	776,25	4,11	23,43	58,94	186,34	-10,6800
39	309,25	469,46	1,97	14,26	36,64	116,20	-1,9173
40	317,34	499,03	2,41	16,09	39,98	130,23	-2,9318
41	350,39	561,72	3,04	17,93	44,81	144,26	-5,5728
42	266,68	384,11	1,45	10,59	26,40	88,15	-0,6463
43	556,88	888,52	4,81	27,10	68,14	214,39	-12,7717
44	587,00	938,54	5,30	28,94	72,75	228,42	-14,6140

Tabla B1: Datos del problema

Bibliografía

Maronna, R.A. and Yohai, V.J. (2014), High finite-sample efficiency and robustness based on distance-constrainedmaximum likelihood. *Computational Statistics and Data Analysis*, **83**, 262–274.

Maronna, R. A., Douglas Martin, R., Yohai, V. J. y Salibián-Barrera, M. (2019), *Robust Statistics, Therory and methods (with R)*, Second edition, John Wiley & Sons.

Maronna, R. A. Regresión: Detalles de las cuestiones básicas para el curso Taller de Estadística. Instituto del Cálculo, Universidad de Buenos Aires. Diciembre 2017.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Romanelli, G.P., Martino, C.M. and Castro, E.A. (2001), Modeling the solubility of aliphatic alcohols via molecular descriptors, Journal of the Chemical Society of Pakistan, Vol.23, No. 4, 195–199.

Yohai, V. J. (1987), High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, vol. 15, no. 2, pp. 642–656.