Taller de Consultoria - TP2 - Selección de Modelos

Gonzalo Barrera Borla

11/10/2019

Introducción

En los tres ejercicios de esta práctica se busca

- (a) elegir un modelo para predecir la variable y a partir de una muestra aleatoria X de un conjunto de covariables $x_1, x_2, \ldots, x_n$, y
- (b) dar una medida del error de predicción asociado al modelo.

A diferencia del TP1, los tres problemas que ahora nos competen ilustran situaciones tan específicas que comprender la relación entre las covariables sin la ayuda de un técnico en el campo es una tarea vana. La *opacidad* del conjunto de covariables (que no es una caja negra, pero tampoco es transparente) nos obliga a intentar ajustar varios modelos en cada situación, y elegir el "mejor" entre ellos, según algún criterio de bondad a definir. Luego, en cada ejercicio intentaremos encontrar el "mejor modelo lineal", siguiendo el mismo procedimiento:

1. **Selección de Modelos**: Compararemos varios modelos posibles entre sí, utilizando como criterio de bondad el ECM de predicción por "validación cruzada deje-uno-fuera" (LOOCV, por sus siglas en inglés). Para evitar el sobreajuste a los datos, utilizaremos la empíricamente eficiente "Regla de 1 Desvío Estándar"[1]. Cconsideraremos que el "mejor modelo" no es el que tenga el menor ECM, sino aquél de menor complejidad (*id est*, con menor cantidad de regresores) con un ECM no más de 1DE por encima del mínimo ECM en términos absolutos.
2. **Análisis de residuos**: Elegido el mejor modelo según la R1DE, revisaremos los residuos de predicción para comprobar que (a) no posean ninguna estructura obvia y (b) sean aproximadamente normales. De violarse severamente alguno de estos supuestos, descartaremos la observación elegida y volveremos al paso 1. De no observar mayores irregularidades, confirmaremos la elección de modelo.

Armados de un criterio sólido de selección de modelos, a continuación hacemos sólo una breve sinopsis de los resultados obtenidos aplicando el mismo criterio a cada problema sin detenernos demasiado en ninguno. El lector avieso notará que los conjuntos de modelos sobre los cuales elegiremos "el mejor" o "el más parsimonioso" distan de ser exhaustivos. Esto es adrede, ya que cualquier enumeración completa de los modelos es una quimera, y estamos realizando este ejercicio casi a modo ilustrativo. En la vida real, sería irresponsable ofrecer un modelo así elegido, sin entender a fondo las covariables involucradas.

Vale la pena recordar que cuando reportemos p-valores y distintos estadísticos de calidad de ajuste para cada modelo, los valores no se pueden tomar literalmente, ya que su interpretación natural ha quedado invalidada por el proceso de selección de modelos y eliminación de observaciones atípicas. A lo sumo, constituyen una medida informal de la bondad del ajuste propuesto, para los datos observados, e la misma manera, el ECM (o si se prefiere, su raíz) provee una medida informal de la precisión del modelo.

Ejercicio 1

La definición de las covariables en el ejercicio tiene algunas particularidades. Por un lado, la "gravedad API" [2] es una escala de densidad específica, propuesta por el American Petroleum Institute. Según las clasificaciones tradicionales, todos los tipos de petróleo evaluados tienen una gravedad mayor a 31.1° (i.e., menor a 870 kg/m3) y por ende son "crudos ligeros". Las otras dos variables con curiosa escala, A y V, están expresadas en "punto ASTM". La "Sociedad Americana para Pruebas y Materiales", o ASTM por sus siglas en inglés, tiene publicados más de 12.000 (!) estándares a la fecha, así que saber a cuál hace referencia la descripción es como buscar una aguja en un pajar. Si fuésemos hombres de apuestas, seguramente nos jugaríamos por el [3] *ASTM D1837-17: Standard Test Method for Volatility of Liquefied Petroleum (LP)*

Gases. Lamentablemente, el estándar fue retirado de circulación en 2017 al haber sido supeditado por métodos más directos y eficientes como la cromatografía de gases.

Como el enunciado pide predecir R en función de G, P, A y V, ignoraremos las variables "tipo" y "corrida", que entendemos son desconocidas al momento de predecir.

Selección de Modelo

A continuación, presentamos en forma de tabla el ECM de LOOCV (y su desvio) para los mejores de entre una selección arbitratria de modelos posibles, que incluye todos los modelos aditivos y multiplicativos de 1 a 4 covariables:

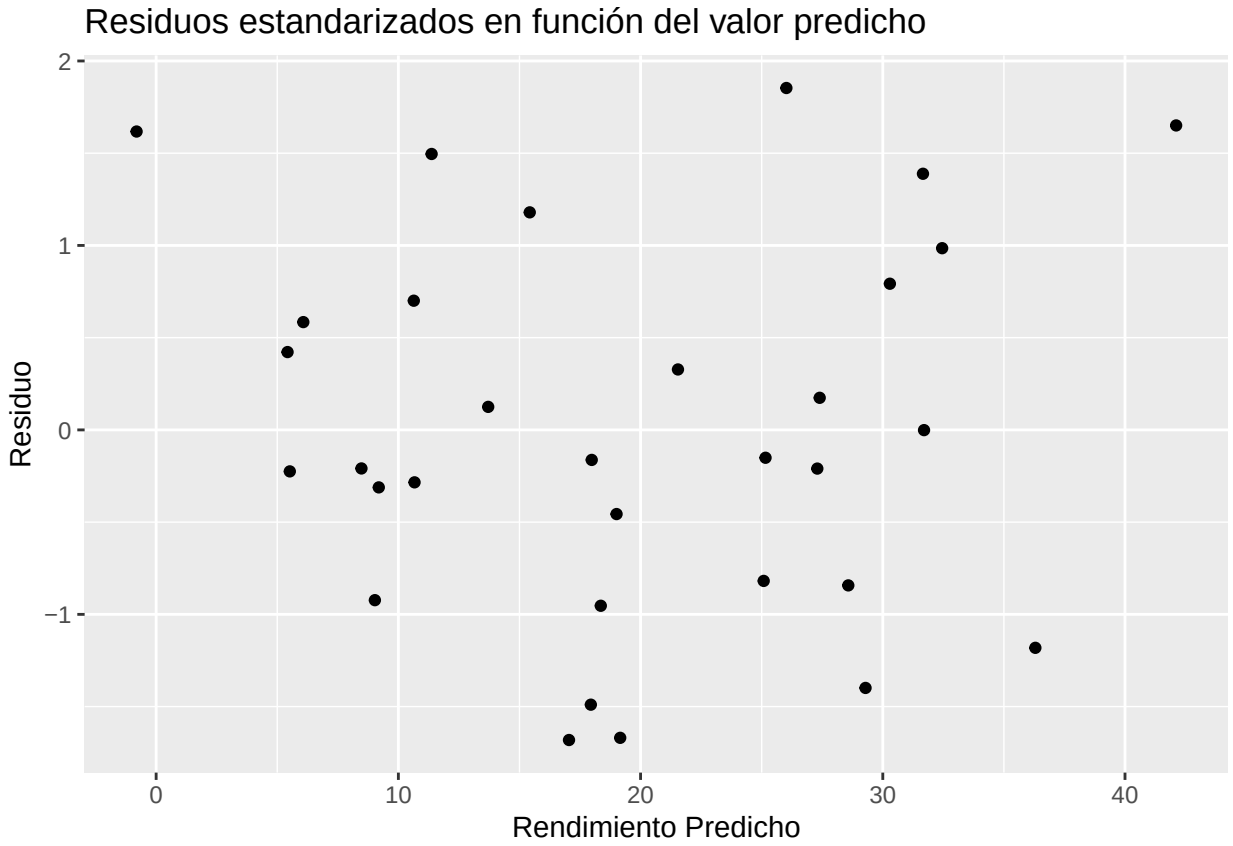
Modelo	ECM	Desvío
R ~ A + P + G + V	5.463	5.816
R ~ A + G + V	5.672	6.244
R ~ A + P + V	6.402	6.821
R ~ A + V	6.511	6.974
R ~ A * G * V	6.637	6.598
R ~ A * V	6.949	7.098
R ~ A * P * V	7.274	7.335
R ~ (A + P + G + V)^2	7.908	7.682
R ~ P * G * V	9.296	9.273
R ~ P + G + V	10.480	11.010
R ~ P * V	13.330	15.080
R ~ A * P * G * V	13.980	19.070

Llamemos ω_{abs} y ω_{1sd} a los modelos con menor ECM en términos absolutos, y al mejor modelo según la R1DE, respectivamente. Resulta entonce que $\omega_{abs} := R \sim A + P + G + V$, con 5 regresores (incluyendo a la ordenada), y dentro de un desvío estándar de su ECM encontramos $\omega_{1sd} := R \sim A + V$, con un ECM un 19% mayor, pero sólo 3 regresores.

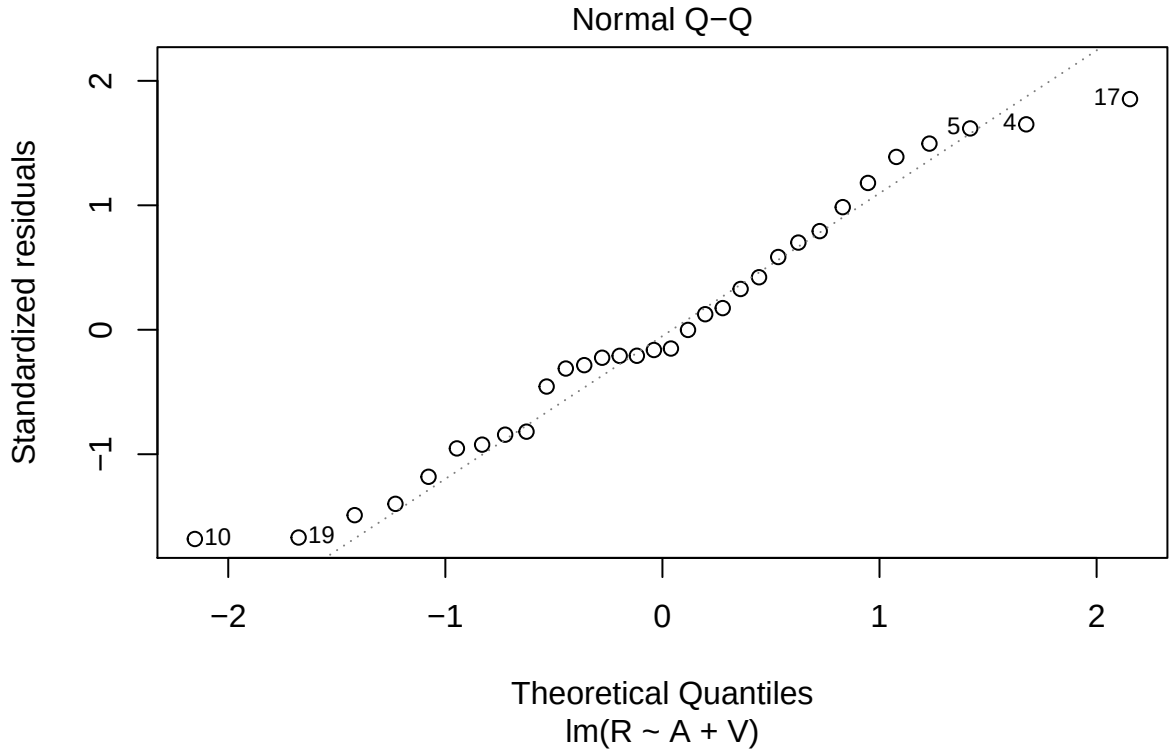
Análisis de Residuos

Para asegurarnos de que el modelo no sólo ajuste bien sino que además esté libre de sesgos sistemáticos, ajustamos ω_{1sd} sobre todos los datos y realizamos

- un *scatterplot* de los residuos estandarizados contra los valores predichos y
- un QQ-plot de los residuos absolutos.



No se observa estructura evidente ni valores extremos de los residuos estandarizados al graficarlos contra predichos. Tampoco hay grandes alejamientos de la normalidad en el QQ-plot, así que confirmamos la selección previa.



Finalmente, reportamos los principales estadísticos del modelo $R \sim A+V$ ajustado *con todas las observaciones* junto con los valores de sus coeficientes:

R^2 aj.	F obs.	P-valor
0.949	288	7.26e-20

Coef.	Estimado	P-valor
(Intercept)	18.5	1.09e-06
A	-0.209	3.11e-16
V	0.156	4.98e-20

Ejercicio 2

Si las variables del ejercicio 1 estaban curiosamente definidas, éstas dan aún más lugar para la imaginación: un "flujo de aire" X_1 en unidades desconocidas, una concentración de ácido X_3 en porcentaje (¿de qué? ¿de los reactivos totales? ¿por qué no se usará el pH?), y una pérdida también en unidades desconocidas. Como especifica el ejercicio, excluimos la variable *dia* de todos los modelos a comparar.

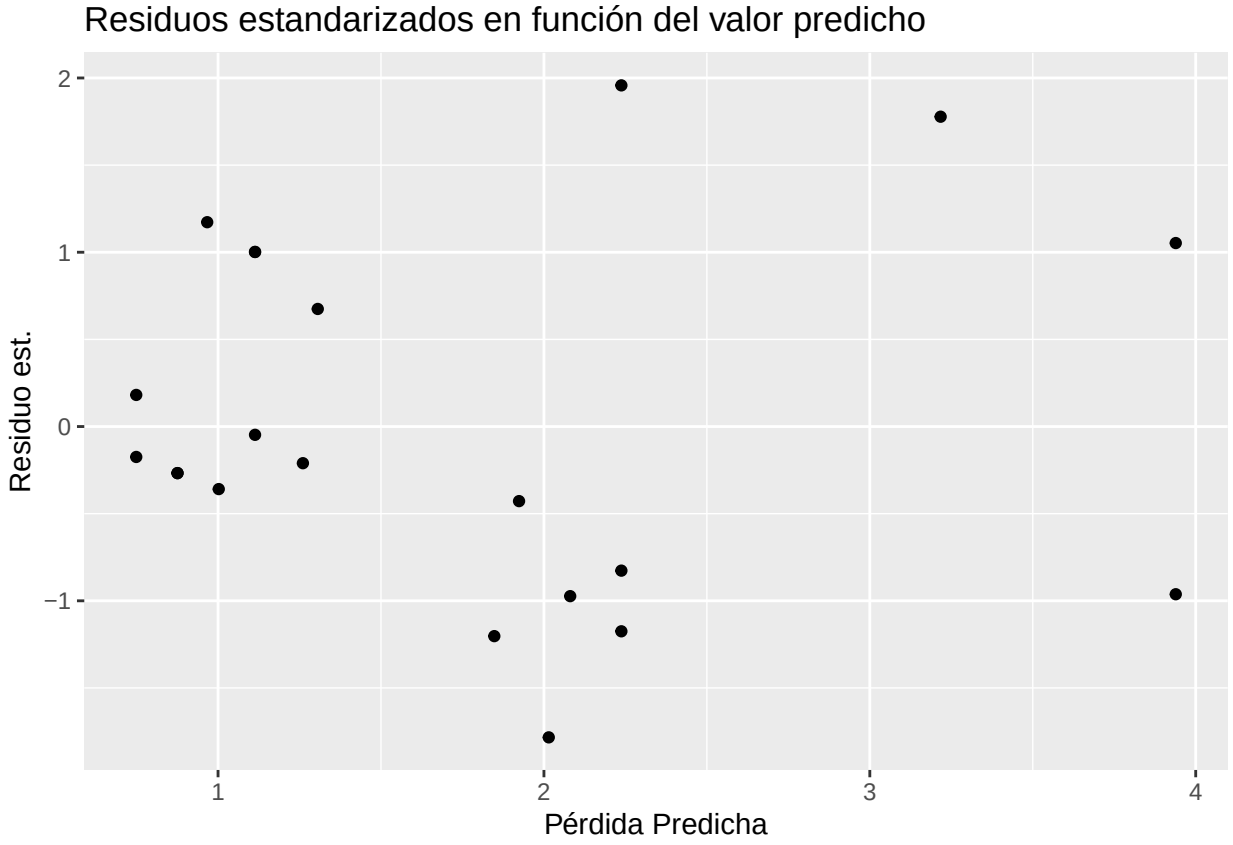
Selección de Modelo

Consideramos los modelos aditivos y multiplicativos para 2 y 3 variables, además de todos los modelos de 1 sola covariable basados en (a) una potencia entera (≤ 3) de una variable original o (b) la interacción de algunas de ellas. Los ECM estimados por LOOCV para los 10 mejores modelos resultan ser:

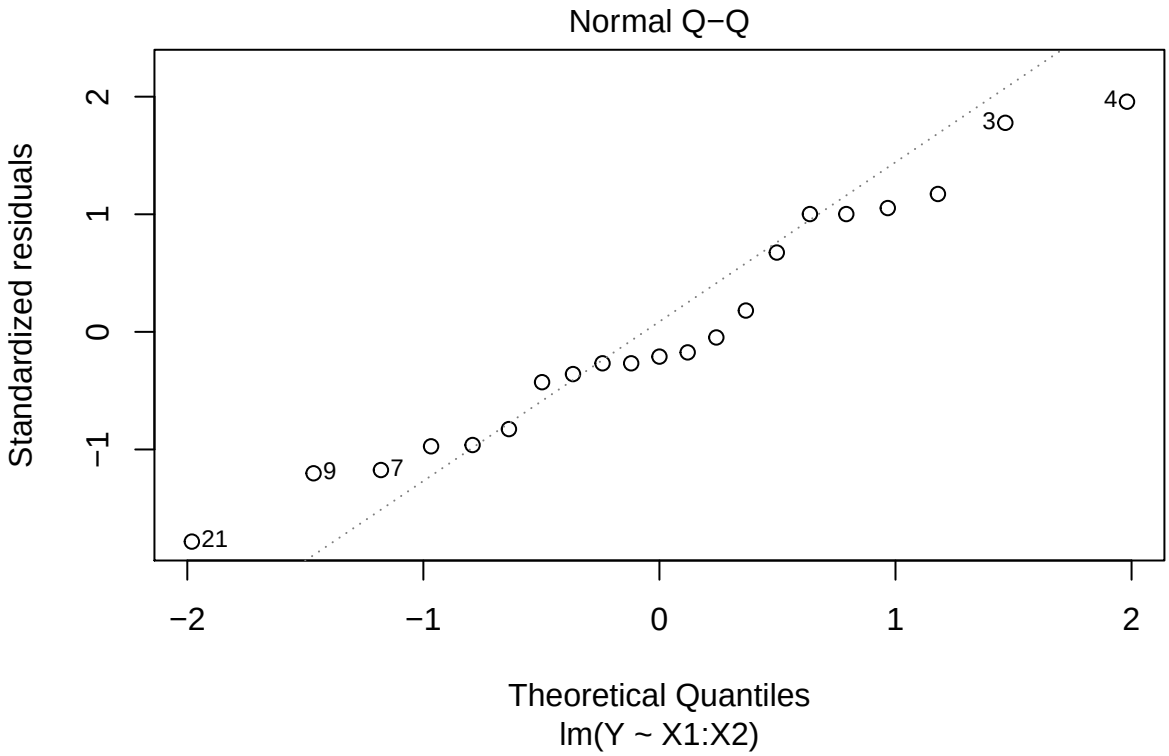
Modelo	ECM	Desvío
Y ~ X1:X2	0.09849	0.1088
Y ~ X1:X2:X3	0.10090	0.1125
Y ~ X1 * X2	0.13570	0.2320
Y ~ X1 + X2 + X3	0.13900	0.2290
Y ~ X1 + X2	0.13980	0.2437
Y ~ X1 * X2 * X3	0.14840	0.2378
Y ~ (X1 + X2 + X3)^3	0.14840	0.2378
Y ~ I(X1^3)	0.17250	0.3959
Y ~ I(X1^2)	0.17470	0.4168
Y ~ (X1 + X2 + X3)^2	0.17680	0.3025

Casualmente, el modelo de menor ECM es también uno de mínima complejidad, con sólo 2 regresores (el modelo de sólo ordenada es el de peor ECM entre los evaluados), así que en principio, $\omega_{abs} \equiv \omega_{1sd} := Y \sim X_1 \times X_2$.

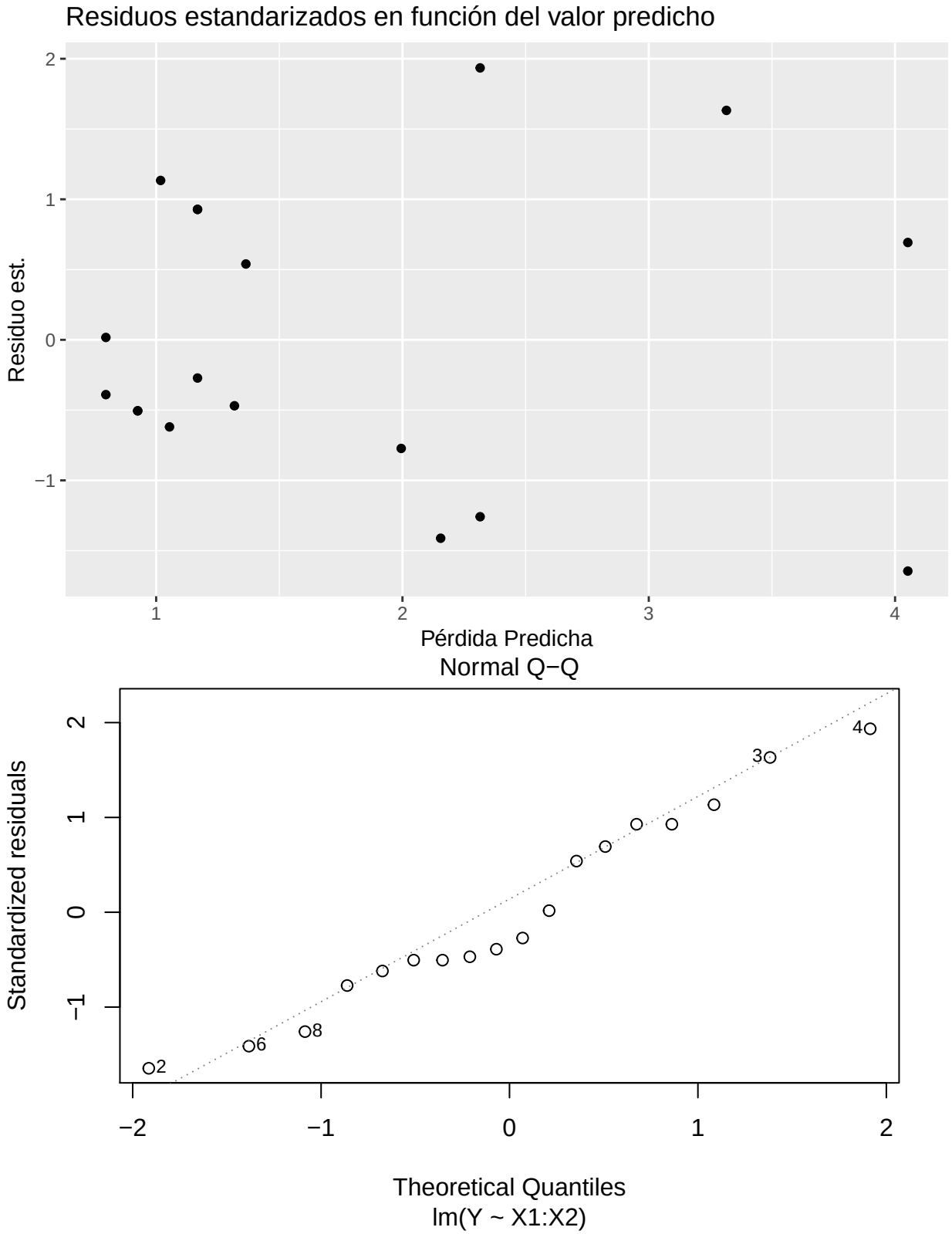
Análisis de residuos



Esta vez, sí observamos cierta estructura en los residuos, pero ninguna tendencia obvia. Si realizamos el mismo gráfico para otros modelos al azar, se nota esta misma seudo-agrupación, así que por el momento no hacemos nada más.



Aquí, observamos cierto alejamiento de la normalidad para los residuos de predicción en algunas observaciones. Para asegurarnos de no haber sesgado la elección de modelo por estos valores extremos, removemos la observación más alejada de la recta cuantil-cuantil (#21) y rankeamos todos los modelos nuevamente por LOOCV. Esto hace que el modelo con el menor ECM absoluto pase a ser $Y \sim X_1 + X_2$, pero $Y \sim X_1 \times X_2$ sigue siendo el mejor modelo según la R1DE. Cuando volvemos a graficar los residuos del modelo ajustado sin la observación 21, vemos que la observación #9 sigue estando bastante alejada de la recta cuantil-cuantil, así que también la descartamos. Repetimos este proceso iterativamente, eliminando primero la observación #9, y luego la #7. En cada caso, $\omega_{abs} := Y \sim X_1 + X_2$, y $\omega_{1sd} := Y \sim X_1 \times X_2$ sigue siendo el mejor modelo según la R1DE. Sin las observaciones $(21, 9, 7)$, el scatterplot de residuos y la recta cuantil-cuantil no muestran mayores irregularidades:



Cuando comparamos tanto los estadísticos globales como los coeficientes particulares del modelo ajustado con y sin las observaciones $(21, 9, 7)$, observamos una sutil mejora en el R^2_{aj}, pero coeficientes iguales hasta el cuarto decimal en ambos casos. Haciendo la misma salvedad de siempre sobre la interpretabilidad de los

p-valores, a continuación reportamos los resultados del modelo ajustado *con todas las observaciones* para el modelo elegido, $Y \sim X_1 \times X_2$:

R^2 aj.	F obs.	P-valor
0.915	217	7.68e-12

Coef.	Estimado	P-valor
(Intercept)	-1.53	2.64e-06
X1:X2	0.00253	7.68e-12

Ejercicio 3

Investigando en la Internet, descubrimos que efectivamente, la espectroscopía del espectro infrarrojo cercano (NIR, por sus siglas en inglés) se utiliza en agricultura [4] para medir la calidad de distintos cultivos y suelos, ya que es un método no-invasivo y relativamente barato. El "estándar dorado", sin embargo, para medir el contenido de nitrógeno en sustancias orgánicas, e indirectamente el contenido proteico de las mismas, es el método de Kjeldahl [5]. Vale aclarar que ambos métodos son formas indirectas de medir la cantidad de proteína en una muestra, y por lo tanto son susceptibles a manipulaciones. En un incidente particularmente cruento [6], en 2008 más de 54,000 bebés fueron hospitalizados y al menos 12 murieron por cálculos renales o insuficiencia proteica cuando varias marcas de leche en polvo adulteraron sus productos con melamina, un químico compuesto en 67% m/m por nitrógeno, más conocido como la materia prima para los revestimientos de fórmica. Esta anécdota ilustra que hacer inferencia sobre cierto soporte de los datos (muestras de calibración) y luego aplicar los resultados obtenidos a muestras por fuera de ese dominio (leche adulterada), no garantiza resultados válidos.

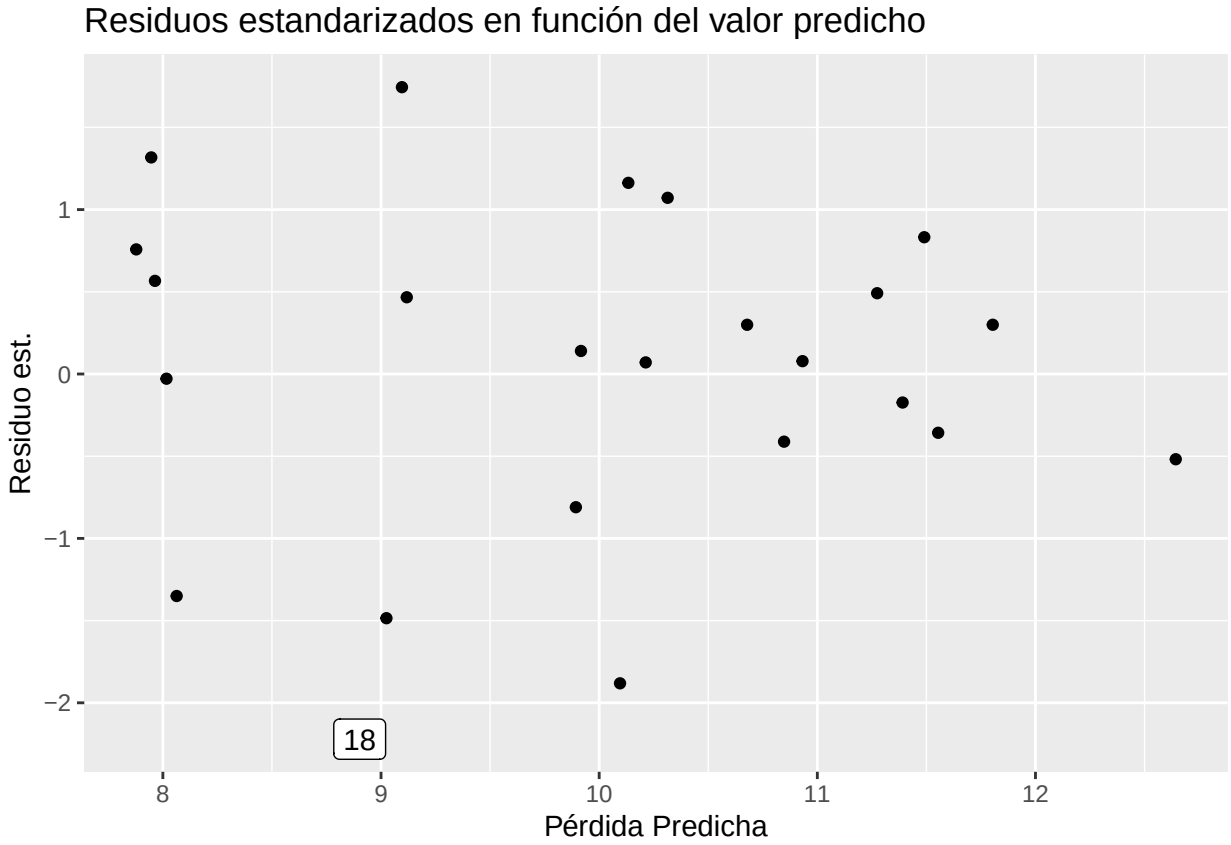
Siendo ésta data de *calibración* de un instrumento, asumimos que las muestras de trigo no están adulteradas, pero tampoco tenemos mucha información sobre la cual trabajar, ya que ni siquiera conocemos las longitudes de onda medidas, como para poder estimar un perfil de absorción de radiación. Así las cosas, con 6 covariables disponibles, sólo los posibles modelos lineales en ellas ascienden a $2^6 = 64$, de modo que *tras bambalinas* primero recurrimos al método `GGally::ggpairs` para graficar los "scatterplots" y calcular la correlación entre cada par de variables disponibles, pero no obtuvimos demasiada información: las seis variables L_i están tan correlacionadas que todos los scatterplots se ven iguales: una línea recta con pendiente positiva. A continuación, sacamos la artillería pesada, y usamos el método `olsrr::ols_all_subset` para ajustar todos los modelos lineales posibles, elegir un subconjunto de los más prometedores, y calcular el ECM según LOOCV como hasta ahora.

Modelo	ECM	Desvío
Y ~ L3 + L4 + L5	0.06573	0.09926
Y ~ L3 + L4 + L5 + L6	0.06584	0.09913
Y ~ L3 + L4	0.06644	0.08746
Y ~ L3 * L4	0.07975	0.10020
Y ~ (L3 + L4)^2	0.07975	0.10020
Y ~ L1 + L2 + L3 + L4 + L5 + L6	0.08122	0.13300
Y ~ L3 * L4 * L5	0.08628	0.13710
Y ~ L3	1.55700	1.74900
Y ~ L3:L4	1.68000	1.76500
Y ~ L3:L4:L5	1.73400	1.77100
Y ~ L3:L5	1.73600	1.82900
Y ~ L4:L5	1.95700	1.91400
Y ~ L4	1.97100	1.89400
Y ~ (L1 + L2 + L3 + L4 + L5 + L6)^2	16.09000	72.35000

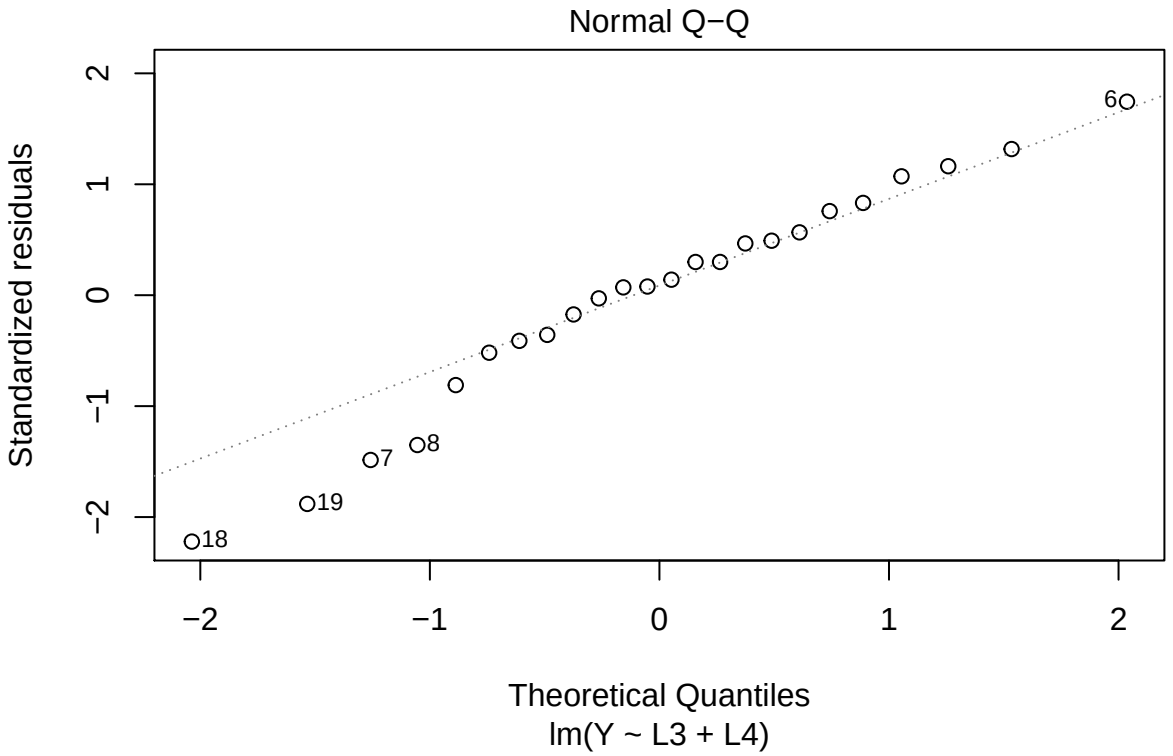
En principio, el mejor modelo "absoluto" es $\omega_{abs} := Y \sim L_3 + L_4 + L_5$, y dentro de 1 ddesvío estándar encontramos $\omega_{1sd} := Y \sim L_3 + L_4$, con 3 regresores. Considerando que el rango de de Y es [3, 12.55], un ECM de 0.066 (RMSE ≈ 0.25) es realmente bueno.

Análisis de Residuos

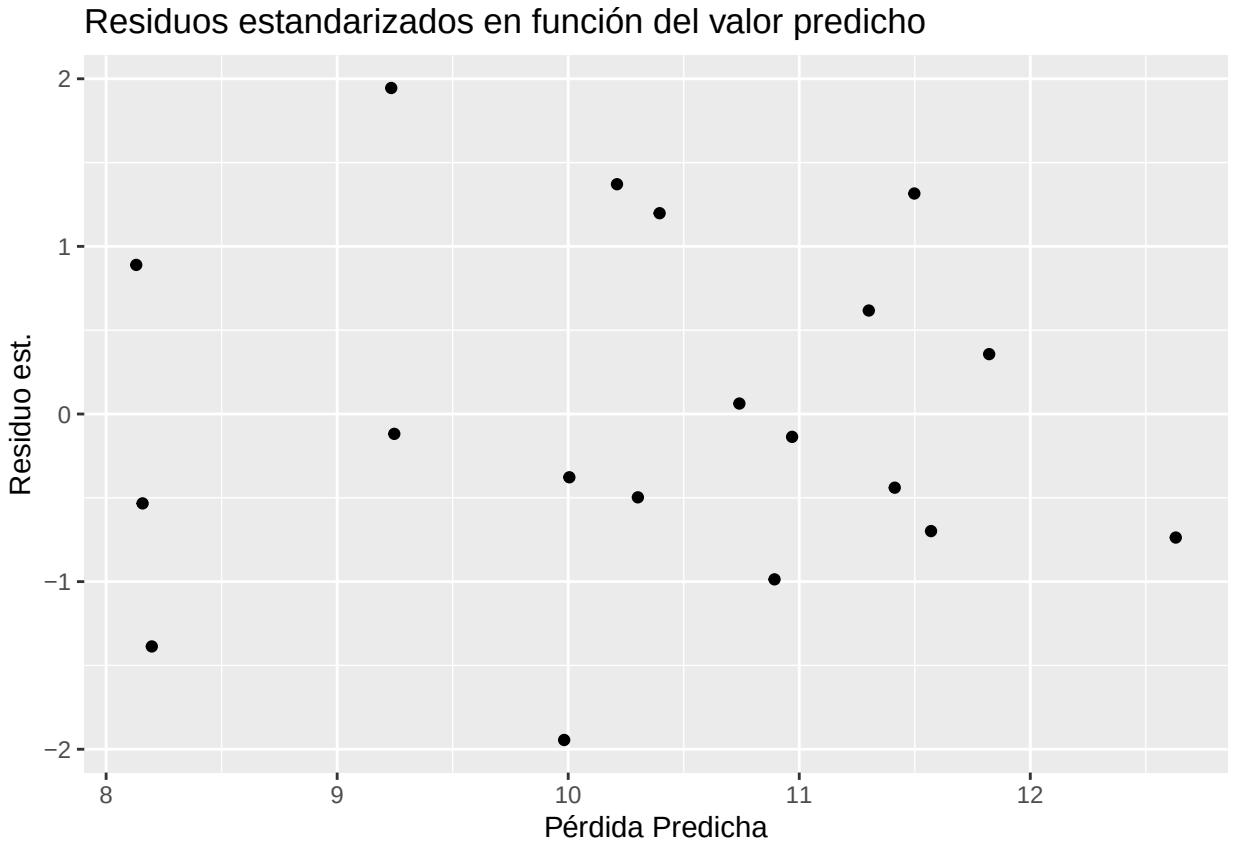
Como siempre, graficamos los residuos para estudiar cualquier estructura remanente.

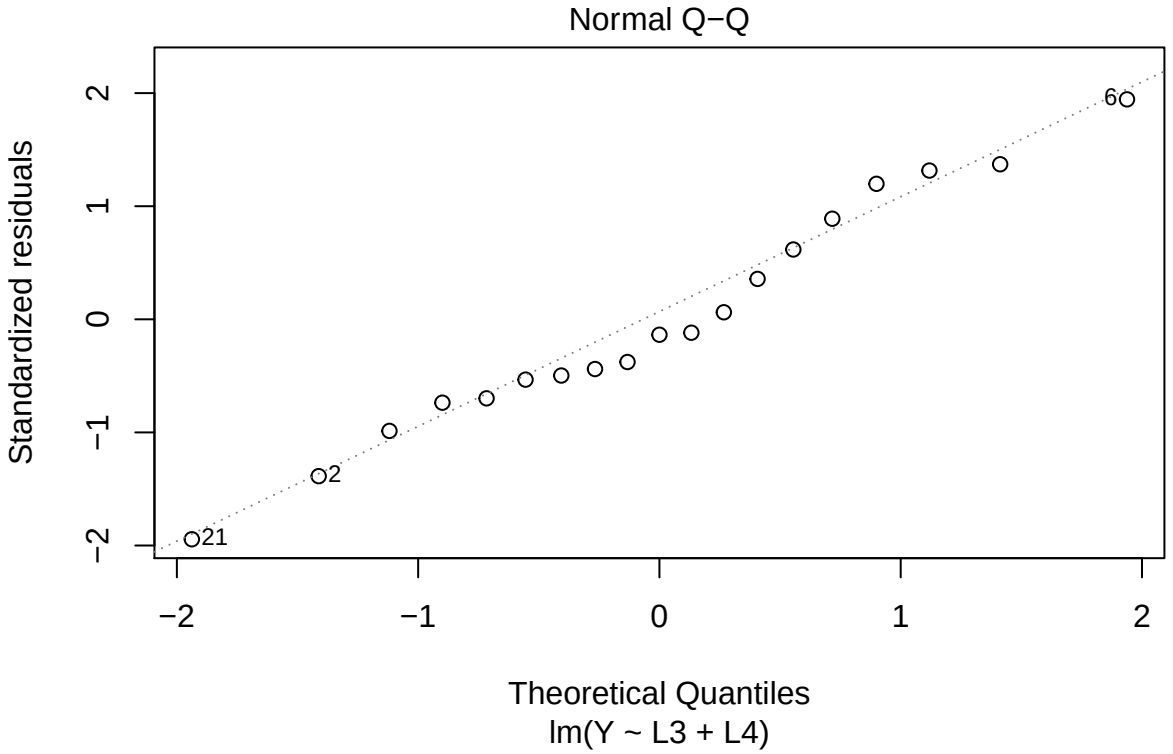


Esta vez, sí observamos un potencial candidato a outlier en la obs. #18, pero ninguna tendencia obvia.



En el QQ-plot de los residuos, resulta evidente que algunas observaciones se alejan considerablemente de los cuantiles esperados. Seguimos el mismo procedimiento de antes de ir eliminando observaciones y recompuntado el ECM para encontrar el mejor modelo según la R1DE, que resulta en la siguiente secuencia de eliminación de observaciones: $(18, 19, 8, 7, 9)$. En cada caso, $Y \sim L_3 + L_4$ resulta ser el mejor modelo tanto en ECM absoluto como según la R1DE siguió siendo $Y \sim L_3 + L_4$, y al eliminarlas tanto los residuos contra predichos como el QQ-plot resultan muy razonables.





Los coeficientes finales no variaron demasiado entre el ajuste con todas las observaciones y sin las "alejadas de la normalidad", con lo cual preferimos conservarlas y reportar estadísticos sobre el modelo ajustado con todas las observaciones:

$$\text{Sin } (18, 19, 8, 7, 9) : \hat{Y} = 30.13 + 0.23L_3 - 0.21L_4$$
$$\text{Con todas} : \hat{Y} = 31.17 + 0.24L_3 - 0.22L_4$$

Al igual que antes, concluimos con los estadísticos principales y coeficientes para el modelo elegido, $Y \sim L_3 + L_4$

R^2 aj.	F obs.	P-valor
0.969	366	4.73e-17

Coef.	Estimado	P-valor
(Intercept)	31.2	1.11e-16
L3	0.24	4.36e-17
L4	-0.217	2.97e-16

Referencias

- [1] Justificacion empírica de la “regla de un desvío estándar” - enlace
- [2] Gravedad API - enlace
- [3] ASTM D1837-17: Standard Test Method for Volatility of Liquefied Petroleum (LP) Gases (Withdrawn 2017) - enlace
- [4] Espectroscopía infrarroja cercana (NIR) en agricultura - enlace
- [5] Método Kjeldahl - enlace
- [6] Adulteración de leche para bebés en 2008 - enlace
- [7] “The Road Not Taken”, de Robert Forst - enlace