Taller de Consultoria - TP2 - Selección de Modelos

Gonzalo Barrera Borla

11/09/2019

Introducción

En los tres ejercicios de esta práctica se busca

- (a) elegir un modelo M 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 a $\hat{y} = M(X)$.

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. Un poco más formalmente,

Sean $M_j : \mathbb{R}^{n \times p} \to \mathbb{R}$ la transformación de X en y que representa al j-ésimo modelo, ω_j. Definamos además la *predicción* $\hat{y}^j = M_j(x_1, \ldots, x_n)$, el *error de predicción* $ep_j = y - \hat{y}^j$ correspondiente a una nueva observación (y_0, x_0) y la complejidad c_j del modelo ω_j. Diremos que ω_k es *más preciso* que ω_l sí y sólo si $ep_k < ep_l$, y que es *más simple* sii $c_k < c_l$.

Si además hacemos el tradicional supuesto de que las observaciones son realizaciones independientes de un mismo proceso generador de datos (DGP, por sus siglas en inglés), es razonable interpretar al error de predicción ep_j como una variable aleatoria a estimar empíricamente. Por conveniencia y tradición, intentaremos estimar el *error cuadrático medio* correspondiente a cada modelo, que no es otra cosa que $\mathbb{E}(ep_j^2)$, con un criterio de "validación cruzada deje-uno-fuera" (LOOCV, en inglés).

Para ello, para cada uno de los m modelos ω_j y las n observaciones x_i, estimaremos n veces los coeficientes de ω_j, *dejando afuera* una observación cada vez, y usaremos el modelo así ajustado para predecir el valor correspondiente a la observación que dejamos afuera. A riesgo de sobrecargar los índices, llamaremos a esta estimación $\hat{y}^{j(i)}$, y los estimadores de la media y varianza de ep_j^2 serán, naturalemnte:

$$\hat{\mu}_j = \widehat{\mathbb{E}(ep_j^2)} = n^{-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{j(i)} \right)^2$$

$$\hat{\sigma}_j^2 = \widehat{\mathbb{V}(ep_j^2)} = (n-1)^{-1} \sum_{i=1}^{n} \left[\left(y_i - \hat{y}_i^{j(i)} \right)^2 - \hat{\mu}_j \right]^2$$

Como toda estimación de una variable aleatoria, $\hat{\mu}_j$ está sujeta a cierta incertidumbre, y tomar como criterio de selección del modelo óptimo $\omega^\star = \{\omega_j : j = \mathrm{argmin}_{j \in \{1,\ldots,m\}} \hat{\mu}_j\}$ no es equivalente a elegir $\omega^\star = \{\omega_j : j = \mathrm{argmin}_{j \in \{1,\ldots,m\}} {ep_j}^2\}$. Si un modelo ω_k es más complejo que otro ω_l ($c_k > c_l$) pero tiene indudablemente menor error de predicción ($ep_k << ep_l$), preferiremos ω_k. Sin embargo, cuando la diferencia entre errores no es tan concluyente, un criterio de parsimonia *alla* "Navaja de Occam" inclinaría la balanza a favor de ω_l. En esta línea argumental, seguiremos la *regla de 1 desvío estándar* y si $\omega^{ep} = \{\omega_j : j = \mathrm{argmin}_{j \in \{1,\ldots,m\}} \hat{\mu}_j\}$ es el modelo que minimiza el estimador de ep^2, el modelo óptimo será el que minimice la complejidad, dentro de los que están a menos de un desvío estándar de $\hat{\mu}_{ep}$,

$$\omega^\star = \{\omega_j : c_j \leq c_{ep} \wedge \hat{\mu}_j \leq \hat{\mu}_{ep} + \hat{\sigma}_{ep}\}$$

Si limitamos la selección de modelos a la familia de *modelos lineales*, Ω_L, una medida razonable de la complejidad de un modelo será la cantidad de coeficientes que ajuste, e irá entre su mínimo en 1 (el modelo basado únicamente en la media global, o los modelos basados en una sola transformación de las covariables y sin ordenada), hasta potencialmente infinito. Nótese además que siempre que $n > p$, $c_j = \#\{\text{coeficientes } \omega_j\} = \text{rg}(\text{H}_j)$, el rango de la matriz de proyección asociada al 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. Cuando sea posible discutiremos la especificidad de las covariables que comprendamos, y justificaremos someramente los modelos a comparar. Ya que encontrar ω^{ep} es prerrequisito para encontrar $\omega^\star$, en general comentaremos sobre ambos modelos un poco más.

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.

Por último, unas palabras sobre la precisión de la predicción. Una métrica razonable y barata de calcular, ya que hemos hecho todo este trabajo y $\hat{\mu}_j$ es el estimador del cuadrado del error, será $\sqrt{\hat{\mu}_j}$, o si se prefiere una tasa, $\sqrt{\hat{\mu}_j}/\bar{y}$. Sin embargo, asumir que estas medidas son sinónimos con el verdadero error de predicción $ep_j = y_0 - \hat{y}_0^j$ o su razón $ep_j/\mathbb{E}(\curvearrowright)$ sería un error. Sin mayor justificación que su intuitividad, usaremos como medida de precisión para e modelo ω_j, el promedio de los desvíos absolutos entre las predicciones de LOOCV y los verdaderos valores, que llamaremos ρ_j:

$$\rho_j = n^{-1} \sum_i \frac{\left| y_i - \hat{y}_i^{j(i)} \right|}{y_i}$$

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.

Una primera dificultad al analizar los datos surge al notar que los pares de covariables $(tipo, P), (tipo, A), (tipo, G)$ son perfectamente colineales. Esto es razonable, ya que P, G, A son *características* de un cierto *tipo* de combustible, pero conducen a matrices de rango incompleto al incluir a más de una de ellas en un modelo. El único tratamiento que se le dio a los datos, entonces, fue

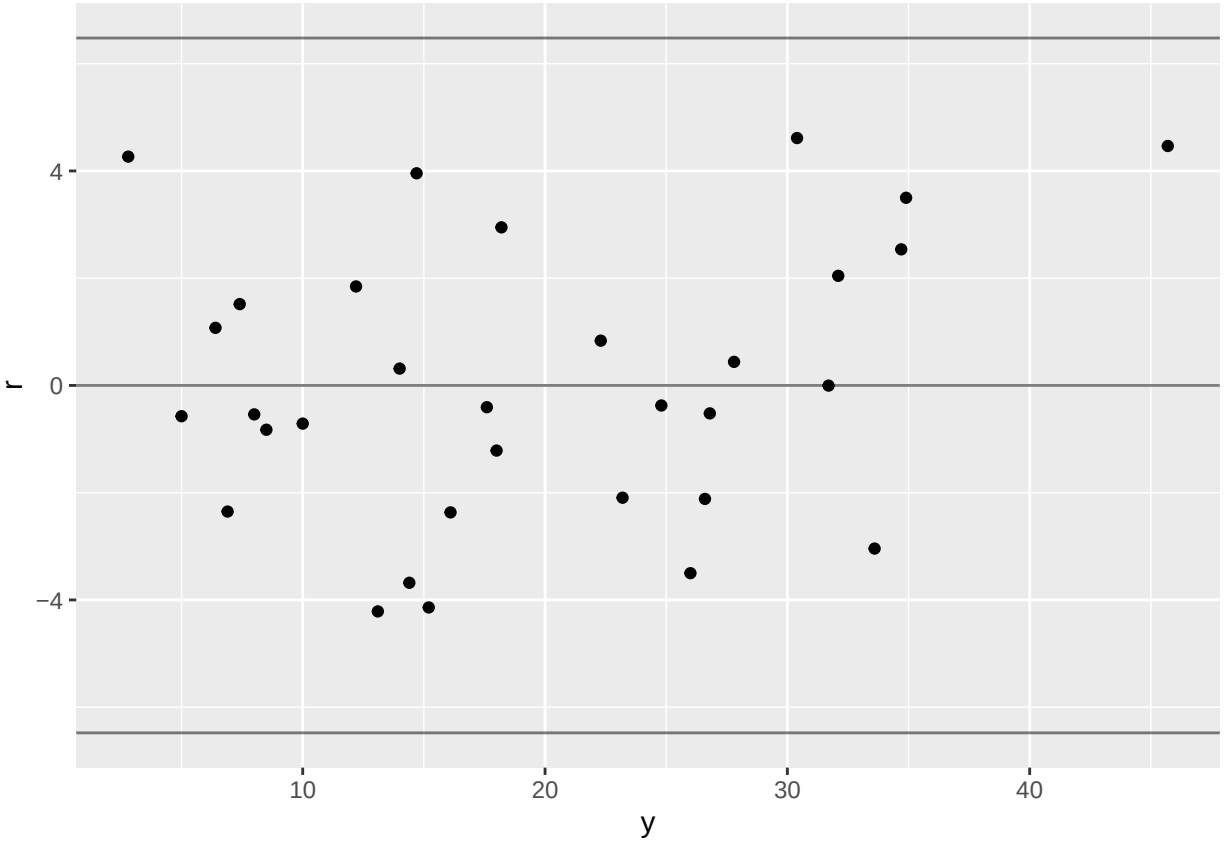
- eliminar la variable *corrida*: el orden de fabricación de cada lote no debería afectar el resultado. Además, hay una fuerte e inexplicable correlación entre la volatilidad final V y la corrida para todos los tipos de crudo, que no conviene incluir.
- transformar la variable *tipo* en factor: si usásemos el valor entero que se provee, estaríamos asumiendo implícitamente un orden entre los distintos tipos de crudo, que la información a mano no justifica. Es preferible tratarla como categórica, y dejar que R genera las "dummies" que hagan falta para representarla.

A continuación, presentamos en forma de tabla las tuplas $(\omega_j,\ \hat{\mu}_j,\ \hat{\sigma}_j, \rho_j)$ para una selección arbitraria de modelos posibles, con 4 cifras significativas:

modelo	ec	desvio	precision
R ~ tipo + A + P + G + V	5.388	7.602	0.1633
R ~ V + tipo	5.388	7.602	0.1633
R ~ A + P + G + V	5.463	5.816	0.1524
R ~ A + G + V	5.672	6.244	0.1481
R ~ corrida + tipo + A + P + G + V	6.003	8.262	0.1736
R ~ A + P + V	6.402	6.821	0.1665
R ~ A + V	6.511	6.974	0.1641
R ~ P + G + V	10.480	11.010	0.1931
R ~ P + V	14.050	13.930	0.2401
R ~ (tipo + A + P + G + V)^2	18.470	59.410	0.2159
R ~ V * tipo	18.470	59.410	0.2159
R ~ G + V	32.040	32.630	0.3940
R ~ V	61.800	82.020	0.4771
R ~ P	109.400	113.800	0.7724
R ~ A	113.200	117.100	0.8035
R ~ A + P	113.700	115.500	0.7796
R ~ P + G	118.300	123.300	0.8026
R ~ G	121.700	130.300	0.8456
R ~ A + G	123.300	128.700	0.8391
R ~ A + P + G	124.200	126.600	0.8150

Resulta entonce que $\omega^{ep} := R \sim tipo + A + P + G + V$, con complejidad $c_{ep} = 6$ (si consideramos a las 9 dummies de *tipo* como una sola) o $c_{ep} = 14$ (si somos estrictos con la definición, pero dentro de un desvío estándar de $\hat{\mu}_{ep}$ encontramos $\omega^\star := R \sim A + V$, con prácticamente la misma precisión (~16%), un error $\hat{\mu}_\star$ muy similar, pero mucha menor complejidad, $c_\star = 3$.

Para asegurarnos de que el modelo no sólo ajuste bien sino que además esté libre de sesgos sistemáticos, graficamos los residuos de las predicciones de LOOCV contra los valores reales. Como referencia, incluimos líneas horizontales en $(-2.5, +2.5)$ desvíos estándar de los residuos muestrales. No se observa ninguna estructura evidente, así que confirmamos la selección previa.



Ejercicio 2

Si las variables del ejercicio 1 estaban curiosamente definidas, estas 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.

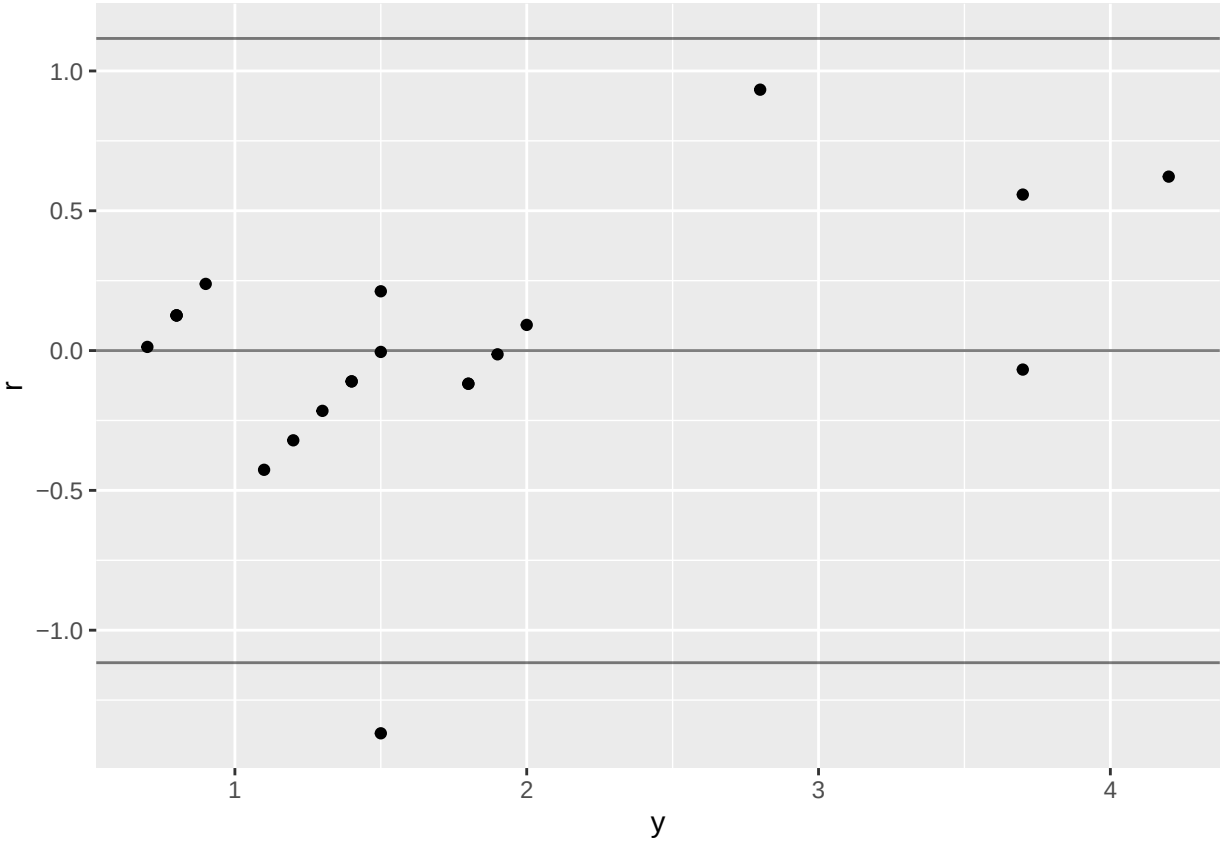
Como especifica el ejercicio, excluimos la variable *dia* de (casi) todos los modelos a comparar, con un par de excepciones por curiosidad intelectual. Presentamos análogo resumen al del ejercicio anterior:

modelo	ec	desvio	precision
Y ~ dia + X1 + X2 + X3	0.1386	0.2078	0.1935
Y ~ X1 + X2 + X3	0.1390	0.2290	0.1845
Y ~ X1 + X2	0.1398	0.2437	0.1617
Y ~ (X1 + X2 + X3)^3	0.1484	0.2378	0.1928
Y ~ (X1 + X2 + X3)^2	0.1768	0.3025	0.1653
Y ~ X1 + X3	0.1887	0.4251	0.1786
Y ~ X1	0.1899	0.4364	0.1727
Y ~ X2	0.2943	0.3122	0.2929
Y ~ X2 + X3	0.3094	0.3249	0.3003
Y ~ X3	0.9520	1.4990	0.4659
Y ~ 1	1.0860	1.7440	0.5348

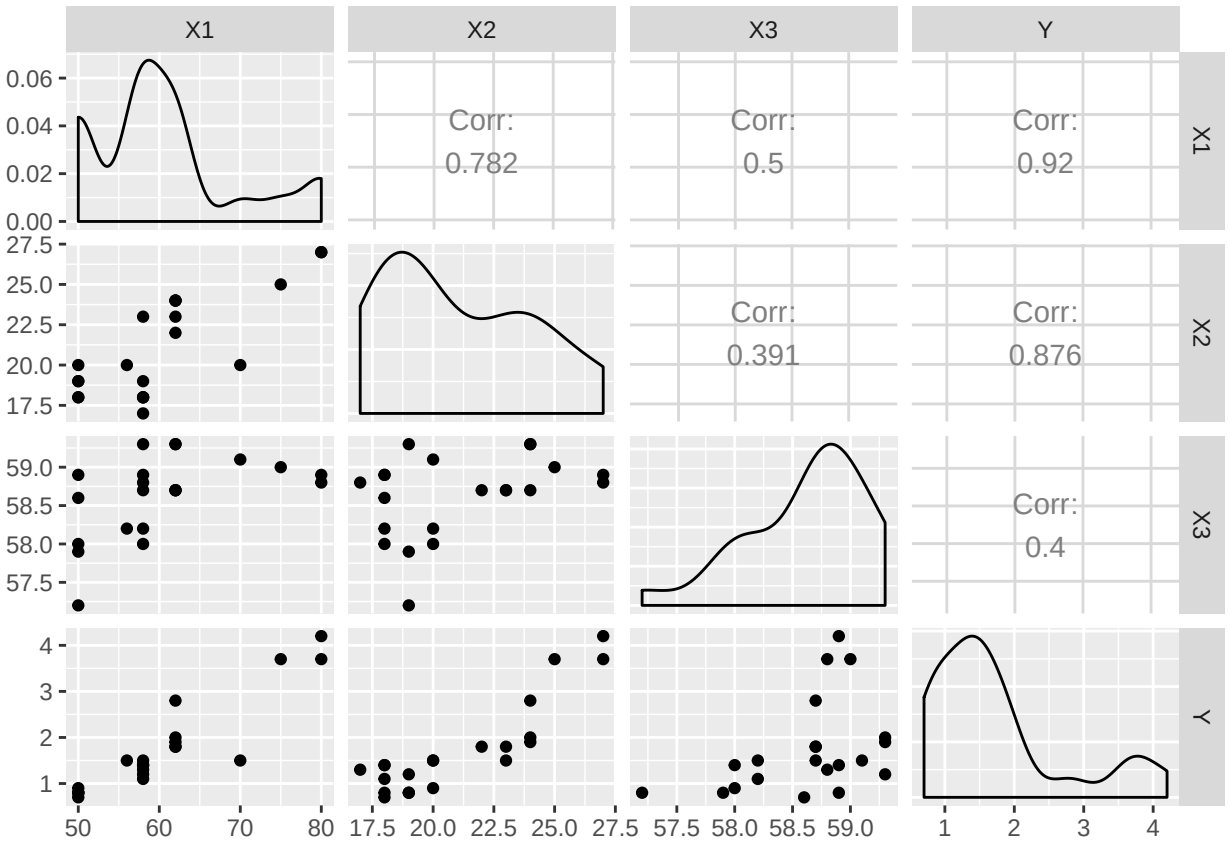
Curiosamente, $\omega_{ep} := Y \sim dia + X_1 + X_2 + X_3$, con un $\hat{\mu}$ ligeramente mejor que el mismo modelo, sin día. En otras palabras, pareciera ser que *algo* de información está codificada en esa secuencia, pero incluirla en el modelo sabiendo tan poco es un tanto irresponsable. En este caso, el desvío estimado es tan alto para ω_{ep} que no sólo uno, sino dos modelos de una sola covariable entran en el intervalo de la R1DE: $Y \sim X_1$ y

$Y \sim X_2$. En principio elegiremos $\omega_\star := Y \sim X_1$ que es (por mucho) el de menor ECM estimado entre ambos. Otro dato interesante, es que aunque $5 = c_{ep} > c_\star = 2$, $\rho_\star \approx 0.17 < 0.19 \approx \rho_{ep}$: ¡la precisión del modelo más simple, asi medida, es mayor que la del modelo más complejo! Esto se debe, probablemente, a que alguna predicción LOOCV de $\omega_\star$ haya sido particularmente mala, lo cual lo penaliza relativamente más en nuestro estimador del ECM $\hat{\mu}$, que en nuestra medida de precisión ρ.

Al igual que antes, graficamos los residuos contra los valores reales para comprobar la existencia de estructura en ellos. Lamentablemente, resulta claro que *algo* extraño hay en los residuos, con al menos un candidato a *outlier* por fuera de la banda de $\pm 2.5SD$, y una intrigante estructura en "zig zag" en los primeros datos.

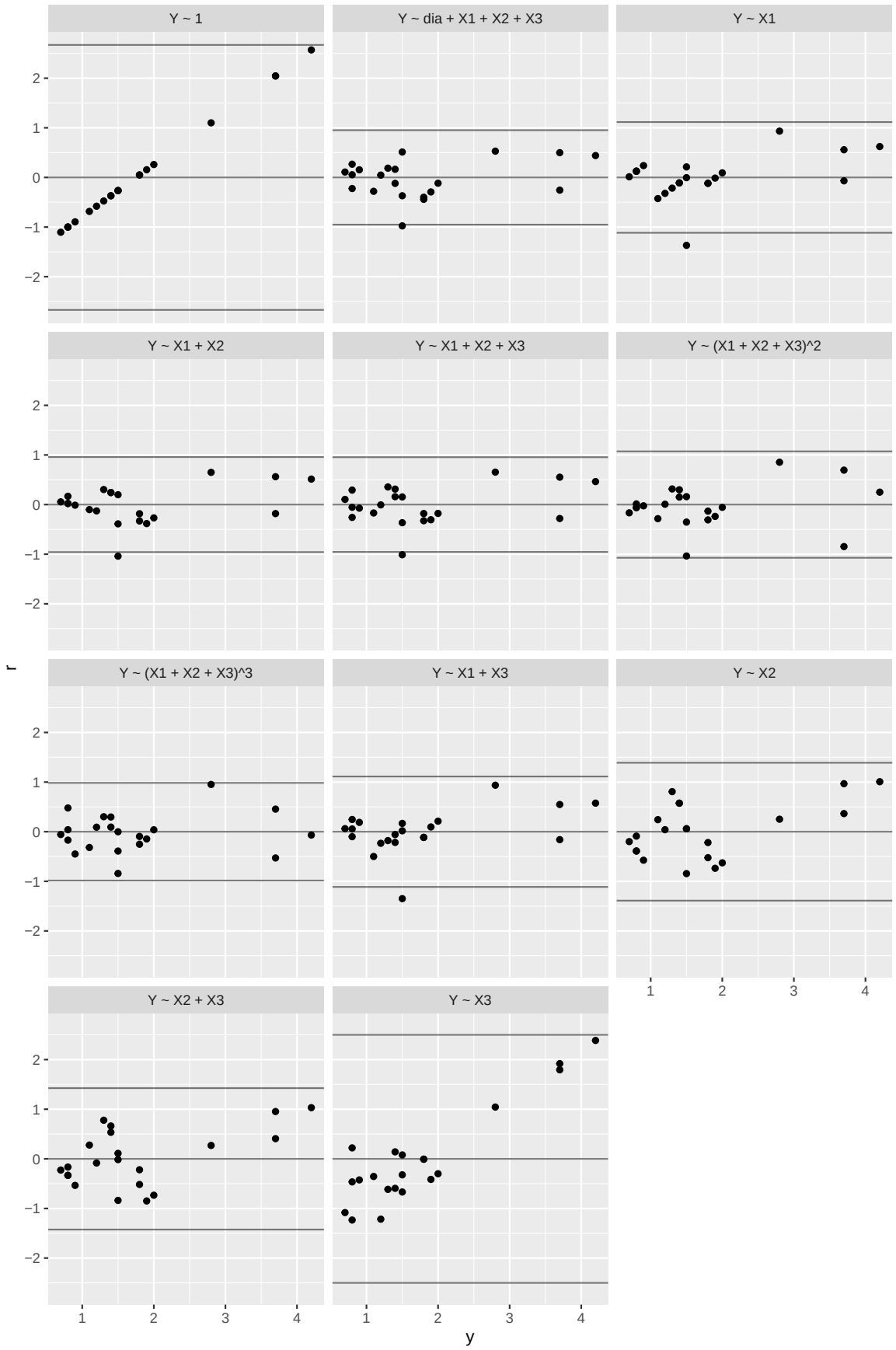


Para entender un poco mejor qué estamos viendo, recurrimos a `GGally::ggpairs`, que realiza una cómoda visualización del los *scatter plots* de cada par de variables de interés, en este caso, (X_1, X_2, X_3, Y):



De aquí, aprendemos dos cosas importantes que explican el gráfico anterior. Primero, que efectivamente hay una observación muy por fuera de la tendencia lineal entre (X_1, Y), y por ello el *outlier* de antes. Y en segundo lugar, se observan muchos valores repetidos de X_1, para distintas Y, que en el modelo $\omega_\star := Y \sim X_1$ darían lugar a predicciones repetidas, lo cual explica las "escaleritas".

De los scatter plots previos, pareciera ser el caso que el *outlier* observado tiene tal carácter por los valores conjuntos de (X_1, Y), y no del resto de las covariables. Para ver si este es el caso, a continuación graficamos los residuos de LOOCV, para *todos* los modelos considerados:


Y ~ 1
Y ~ dia + X1 + X2 + X3
Y ~ X1
Y ~ X1 + X2
Y ~ X1 + X2 + X3
Y ~ (X1 + X2 + X3)^2
Y ~ (X1 + X2 + X3)^3
Y ~ X1 + X3
Y ~ X2
Y ~ X2 + X3
Y ~ X3
r
y

Efectivamente, vemos que cada vez que incluimos a X_1 en un modelo, nos topamos con exactamente una predicción descabellada, y si no la incluimos, no se observa tal outlier. Aquí, dos caminos divergen en el "bosque" del análisis [7]. O bien

- consideramos efectivamente como *outlier* a la observación en duda, rehacemos el análisis sin ella, el ECM de todos los modelos que incluyan a X_1 mejora, y casi seguramente elijamos los mismos ω_{ep}, $\omega_\star$ que antes, o
- **no** consideramos *outlier* al dato, lo mantenemos dentro del *dataset* de entrenamiento, y corregimos la selección de modelos para asegurarnos que los residuos del modelo elegido no presenten estructura obvia.

En la vida real, nos acercaríamos al técnico o analista que nos haya provisto los datos y le preguntaríamos por la observación anómala, para saber si conservarla tal cual, corregirla o descartarla. Como no podemos darnos ese lujo, debemos tomar una decisión "a ciegas", y siguiendo un critero de "observación inocente (válida) hasta que se pruebe lo contrario", preferimos conservarla y **cambiar la selección de modelo a** $\omega_{\star'} := Y \sim X_2$, que cumple la condición de parsimonia (R1DE) y no presenta mayores particularidades en la estructura de los residuos.

Por último, y anecdóticamente, nótese que los modelos cuadrático ($c_2 = 10$) y cúbico ($c_3 = 20$, si no me equivoco) en todas las covariables no agregan nada de información, y por ende no mejoran el ECM. Tal vez, si contásemos con miles y miles de observaciones de las cuales extraer estructura y las mismas únicas 3 covariables, la historia sería distinta.

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 orgnánicas, 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. Como bien hemos aprendido ya, hacer inferencia sobre cierto soporte de los datos y luego aplicar los resultados obtenidos a muestras por fuera de ese dominio, no garantiza resultados válidos.

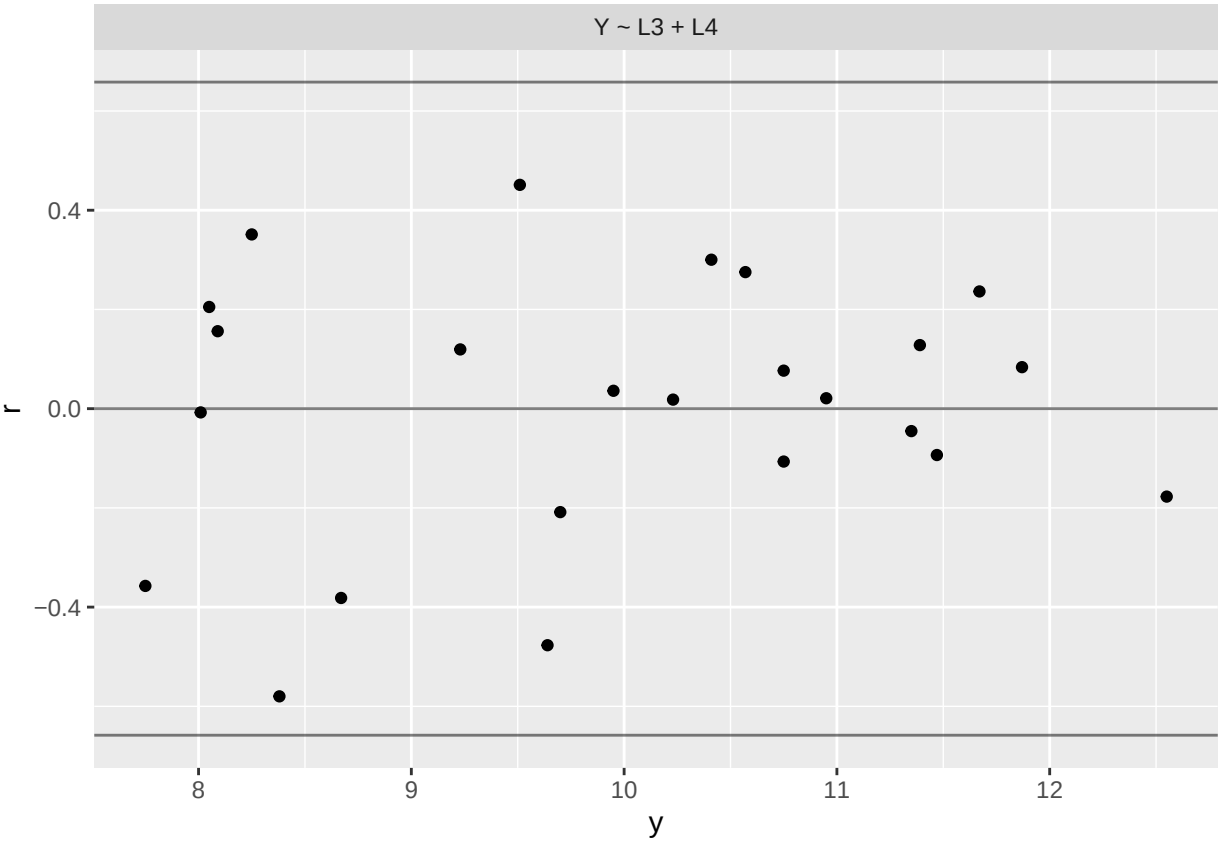
Siendo ésta data de *calibración* de un instrumento, asumimos que las muestras de trigo no están adulteradas, pero tampoco tenemos mucha información sobre la cual trabajar, ya que ni siquiera conocemos las longitudes de onda medidas, como para poder estimar un perfil de absorción de radiación. Así las 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	ec	desvio	precision
Y ~ L3 + L4 + L5	0.06573	0.09926	0.02160
Y ~ L3 + L4 + L5 + L6	0.06584	0.09913	0.02063
Y ~ L3 + L4	0.06644	0.08746	0.02179
Y ~ (L3 + L4)^2	0.07975	0.10020	0.02347
Y ~ L1 + L2 + L3 + L4 + L5 + L6	0.08122	0.13300	0.02253
Y ~ muestra + L1 + L2 + L3 + L4 + L5 + L6	0.08336	0.13310	0.02323
Y ~ L3	1.55700	1.74900	0.10940

modelo	ec	desvio	precision
Y ~ L4	1.97100	1.89400	0.12590
Y ~ (L1 + L2 + L3 + L4 + L5 + L6)^2	16.09000	72.35000	0.15070

$\omega_{ep} := Y \sim L_3 + L_4 + L_5$, con $c_{ep} = 4$, y $\omega_\star := Y \sim L_3 + L_4$, con $c_\star = 3$. La precisión de ambos modelos es muy similar y dado lo indirecto del método, sorprendentemente buena: ligeramente por encima del 2%, comparado con el ~16% de los ejercicios previos.

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



En principio, no observamos una tendencia obvia en los residuos. Podríamos llegar a sospechar cierta heterocedasticidad, con menor varianza a medida que Y aumenta, pero la cantidad de observaciones disponibles es más bien pobre para concluir al respecto. En principio, confirmamos la selección de modelo, dejando un "asterisco" en el supuesto de homocedasticidad para testear su validez en caso de obtener más datos.

Como en los ejercicios previos, si trazásemos un intervalo de confianza alrededor del estimador puntual, $0 \in (\hat{\mu}_{ep} \pm 1\hat{\sigma}_{ep})$. Evidentemente, (y esto es natural considerando que el ECM es siempre positivo), la distribución no es simétrica alrededor de la media, y ofrecer un verdadero IC para el ECM no es ejercicio trivial. En la misma línea, tal vez a causa del pequeño tamaño muestral $\hat{\mu}_{ep} < \hat{\sigma}_{ep}$, y la regla de 1SD deja verdaderamente mucho espacio para elegir modelos parsimoniosos. En este caso, resulta ilustrativo ver que por separado $Y \sim L_3$ y $Y \sim L_4$ son modelos con enorme error, pero el modelo combinado es ~6 veces más preciso y con ~25 veces menor ECM.

Por último, se ve que con variables tan correlacionadas, agregar los términos cuadráticos al modelo dispara el ECM de VC al demonio (aunque de seguro no puede disparar el ECM de entrenamiento), cosa que no necesariamente sucede cuando hay menos términos disponibles como en el ejercicio 2.

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