Práctica 8: Regresión lineal múltiple

En esta práctica nuestro objetivo es realizar un análisis de regresión lineal múltiple con un conjunto de datos. Para ello utilizaremos la función `lm` junto con otras funciones relacionadas que nos permite extender los resultados de la misma.

1. Contexto: Datos de marketing

El contexto del análisis es un estudio de mercado donde se obtuvieron datos de 100 empresas, clientes de un gran distribuidor industrial (denominado HATCO). Los datos se recogen en el fichero `hatco2.txt`, disponible en PRADO, que contiene para cada empresa información acerca de las siguientes variables:

Variable	Descripción
`empresa`	Identificador de la empresa
`tamano`	Tamaño de la empresa
`adquisic`	Estructura de adquisición
`tindustr`	Tipo de industria
`tsitcomp`	Tipo de situación de compra
`velocidad`	Velocidad de entrega
`precio`	Nivel de precios
`flexprec`	Flexibilidad de precios
`imgfabri`	Imagen del fabricante
`servconj`	Servicio conjunto
`imgfvent`	Imagen de fuerza de ventas
`calidadp`	Calidad de producto
`fidelidad`	Porcentaje de compra a HATCO
`nfidelidad`	Nivel de compra a HATCO
`nsatisfac`	Nivel de satisfacción

Inspecciona el fichero (desde por ejemplo el bloc de notas) para comprobar la estructura que tiene del fichero y el tipo de datos. Teniendo esto en cuenta carga después los datos en R, almacenándolos en un data frame con nombre `hatco` donde las variables de tipo factor se identifiquen como tal.

De las 16 variables que componen el data frame `hatco` que has creado 9 deben ser numéricas[1] y el resto factores. Entre las numéricas, las variables `velocidad`, `precio`, `flexprec`, `imgfabri`, `servconj`, `imgfvent` y `calidadp`, constituyen percepciones del cliente (la empresa) acerca de la distribuidora y sus productos en relación a distintos aspectos. Estas

[1] La variable `empresa` aunque es numérica constituye un identificador de la empresa.

percepciones han sido evaluadas en una escala métrica entre 0 (pobre) y 10 (excelente). La variable `fidelidad` se mide como el porcentaje que se compra al distribuidor del total del producto de la empresa.

2. Análisis de regresión lineal múltiple

2.1. Objetivo y modelo

Nuestro objetivo con los datos anteriores es predecir los niveles de fidelidad al distribuidor por parte de los clientes, tomando con base las percepciones que estos tienen del mismo y de sus productos, así como identificar los factores que llevan al aumento de la utilización del producto. Para ello se propone un modelo de regresión lineal múltiple:

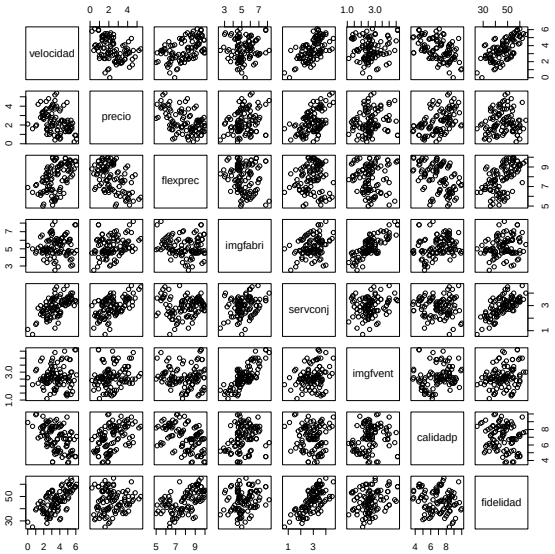
$$\begin{aligned} Y_i &= \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \epsilon_i \quad (i = 1, \ldots, n) \\ &\quad \text{asumiendo que } \epsilon_i \rightsquigarrow N(0, \sigma^2) \text{ (incorreladas)} \end{aligned}$$

donde la variable de respuesta Y es la fidelidad por parte del cliente (`fidelidad`), y como variables explicativas se consideran las $k = 7$ percepciones medidas: `velocidad`, `precio`, `flexprec`, `imgfabri`, `servconj`, `imgfvent` y `calidadp`.

2.2. Representación gráfica y ajuste del modelo

Antes de ajustar el modelo anterior representamos gráficamente los datos para observar la posible relación entre la respuesta y las variables explicativas.

```
> plot(hatco[,c(6:13)])
```



El gráfico obtenido proporciona diagramas de dispersión de las variables consideradas dos a dos. De este gráfico, y siempre de una manera aproximada, es posible:

- Advertir si existe relación lineal entre la respuesta y cada una de las variables explicativas.
- Descubrir si hay variables explicativas que sean aproximadamente colineales al resto.

Observa el gráfico en relación a los dos puntos anteriores, ¿qué podrías decir?

Ajustamos ahora el modelo de regresión lineal múltiple a las $n = 100$ observaciones que tenemos de las variables, usando la función `lm`, y almacenando el resultado en un objeto con nombre `mod1`:

```
> mod1<-lm(fidelidad~velocidad+precio+flexprec+imgfabri+servconj+
+            imgfvent+calidadp,hatco)
> mod1


Call:
lm(formula = fidelidad ~ velocidad + precio + flexprec + imgfabri +
    servconj + imgfvent + calidadp, data = hatco)

Coefficients:
(Intercept)    velocidad       precio     flexprec     imgfabri     servconj
  -10.16148     -0.04352     -0.67891      3.36197     -0.04101      8.34537
   imgfvent     calidadp
    1.29147      0.56295
```

El resultado nos muestra los coeficientes estimados, $\widehat{\beta}_0, \ldots, \widehat{\beta}_7$, de donde los valores ajustados (estimados) de la fidelidad para cada cliente se obtienen a partir de la siguiente expresión lineal:

$$\widehat{Y}_i = -10.161 - 0.044x_{i1} - 0.679x_{i2} + 3.362x_{i3} - 0.041x_{i4} + 8.345x_{i5} + 1.291x_{i6} + 0.563x_{i7}$$

Los coeficientes estimados $\widehat{\beta}_j$ representan la magnitud del efecto que cada percepción del cliente ejerce en su fidelidad a la distribuidora. Por ejemplo, considerando la flexibilidad de precios (x_3) podemos decir que por cada unidad que aumenta la percepción que el cliente tiene de dicha flexibilidad, su fidelidad se incrementa en 3.362 unidades, supuesto que el resto de percepciones permanece constante.

2.3. Inferencia sobre el modelo

Bondad del ajuste y contraste de regresión

Estudiamos ahora en qué medida las 7 percepciones de forma conjunta consiguen describir la fidelidad de los clientes y hacemos una valoración global del ajuste. Para ello

calculamos el contraste de regresión y los coeficientes R^2 (coeficiente de determinación) y $\bar{R}^2$ (versión corregida para la regresión múltiple). Todo esto lo podemos obtener con la función `summary`.

```
> summary(mod1)


Call:
lm(formula = fidelidad ~ velocidad + precio + flexprec + imgfabri +
    servconj + imgfvent + calidadp, data = hatco)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9759  -1.9491   0.5896   2.8144   6.7565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.16148    4.97696  -2.042   0.0440 *
velocidad    -0.04352    2.01273  -0.022   0.9828
precio       -0.67891    2.09025  -0.325   0.7461
flexprec      3.36197    0.41125   8.175 1.56e-12 ***
imgfabri     -0.04101    0.66683  -0.061   0.9511
servconj      8.34537    3.91830   2.130   0.0359 *
imgfvent      1.29147    0.94720   1.363   0.1761
calidadp      0.56295    0.35544   1.584   0.1167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.424 on 92 degrees of freedom
Multiple R-squared:  0.775,Adjusted R-squared:  0.7578
F-statistic: 45.26 on 7 and 92 DF,  p-value: < 2.2e-16
```

El contraste de regresión se muestra al final de la salida anterior. Este contraste formula la hipótesis nula $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$, lo que implicaría que las variables explicativas de forma conjunta no tienen ningún efecto en la variable de respuesta. Esto parece poco probable atendiendo a la naturaleza de las variables, y se confirma con el estadístico de contraste obtenido $F = 45.26$ y el p-valor asociado de aproximadamente 0. Por tanto podemos concluir que el modelo de regresión ajustado es útil y tiene sentido, ya que las variables explicativas (percepciones de los clientes en este caso) consideradas de modo conjunto permiten explicar la variable de respuesta (fidelidad).
En relación a la bondad del ajuste observamos los valores `Multiple R-squared` y `Adjusted R-squared`. El primero es el coeficiente de determinación, R^2, que en este caso vale 0.775). Lo podemos interpretar diciendo que aproximadamente el 77.5 % de la variabilidad total

de la fidelidad de los clientes queda explicada por las 7 percepciones a través del modelo lineal ajustado. El segundo de los valores es el coeficiente de determinación corregido que resulta 0.7578. Este valor da una medida adecuada[2] de la bondad del ajuste, corregido por el número de variables explicativas.

Tabla ANOVA

La tabla ANOVA nos muestra la descomposición de la variabilidad utilizada para el contraste de regresión descrito antes y el coeficiente R^2. La obtenemos usando la función `anova`:

```
> anova(mod1)

Analysis of Variance Table

Response: fidelidad
          Df Sum Sq Mean Sq  F value    Pr(>F)
velocidad  1 3659.8  3659.8 187.0071 < 2.2e-16 ***
precio     1  927.9   927.9  47.4128 6.932e-10 ***
flexprec   1 1346.3  1346.3  68.7912 8.779e-13 ***
imgfabri   1  100.4   100.4   5.1311   0.02585 *
servconj   1   71.5    71.5   3.6517   0.05913 .
imgfvent   1   44.9    44.9   2.2963   0.13311
calidadp   1   49.1    49.1   2.5084   0.11667
Residuals 92 1800.5    19.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La columna `Sum Sq` nos da al final la variabilidad no explicada (`Residuals`) y en las filas anteriores la variabilidad explicada de forma secuencial para cada variable explicativa (sumándolas tendríamos la variabilidad explicada por el modelo ajustado).

Significación individual de las variables explicativas

Estudiamos ahora la influencia individual que cada una de las 7 percepciones consideradas tiene en la fidelidad. Para ello formulamos contrastes de significación individuales para cada percepción x_j $(j = 1, \ldots, 7)$, de tipo $H_0 : \beta_j = 0$ frente a $H_1 : \beta_j \neq 0$. Se trata por tanto de 7 problemas de contraste de hipótesis donde en cada caso rechazar la hipótesis nula supondrá concluir que la percepción correspondiente tiene influencia significativa en la fidelidad, considerada en el contexto de la regresión lineal múltiple (donde están presentes las 7 percepciones). Los resultados de cada contraste se mostraron anteriormente

[2] El R^2 tiende a sobrevalorar el ajuste tomando valores mayores a medida que incluímos variables explicativas adicionales, aunque no tengan ninguna relación con la respuesta.

como resultado aplicar la función `summary` (observa los contrastes en la tabla de los coeficientes que entre otra información te ofrecía dicha función antes).

Por ejemplo para la primera variable explicativa, x_1, correspondiente a la percepción que el cliente tiene de la velocidad de entrega, el estadístico de contraste[3] sería: $t_1 = \widehat{\beta}_1/s.e.(\widehat{\beta}_1) = -0.0435/2.0127 = -0.0216$. Observando el p-valor asociado, 0.983, se concluye que no hay evidencia suficiente de los datos para rechazar H_0 al 5 % de significación, y por tanto la percepción que el cliente tiene de la velocidad de la entrega no parece influir significativamente en su fidelidad a la distribuidora[4]

Repite el razonamiento anterior para las otras percepciones $x_2, \ldots, x_7$. ¿Qué puedes decir de la significación individual del resto de percepciones al 5 %.

Significación del término constante β_0

Además de los efectos de cada percepción β_j $(j = 1, \ldots, 7)$, el modelo que hemos ajustado incluye un término constante, el cual se interpreta en general como el valor medio de la respuesta cuando todas las variables explicativas toman el valor 0. En el ajuste realizado la estimación de este término es $\widehat{\beta}_0 = -10.1615$ con un error estándar de 4.977. Podemos estudiar su significación formulando el problema de contraste $H_0 : \beta_0 = 0$ frente a $H_1 : \beta_0 \neq 0$. Los resultados están de nuevo en la tabla de coeficientes que proporcionaba la función `summary` (fila correspondiente a `Intercept`).

Considerando un nivel de significación del 5 %, ¿se podría prescindir del término constante en el modelo? ¿Y al 1 %?

2.4. Diagnósticos del modelo

A continuación verificamos las hipótesis del modelo de regresión que hemos considerado inicialmente.
Los diagnósticos del modelo incluyen en primer lugar un análisis de los residuos para comprobar que se verifican las hipótesis del modelo de regresión lineal múltiple: linealidad de la relación y normalidad, homocedasticidad e incorrelación de los errores del modelo. A través de este análisis también comprobamos si existen observaciones anómalas e influyentes[5]. Finalmente se hará un estudio de la posible multicolinealidad entre las variables explicativas del modelo.

[3] Bajo H_0 el estadístico de contraste sigue una t de Student con $n - k - 1$ grados de libertad (en este caso serían 92 grados de libertad).

[4] Esta conclusión sin embargo debe entenderse en el contexto del modelo ajustado, donde están presentes las 7 percepciones, y no nos lleva a concluir que podemos prescindir de x_1 en el modelo, sino que es posible que el modelo se pueda simplificar ya que alguna(s) variable(s) puede(n) ser redundante(s).

[5] Algunos gráficos para estos diagnósticos se pueden obtener escribiendo `plot(mod1)`. Puedes echar un vistazo a los mismos no obstante en esta práctica construiremos nuestros propios gráficos siguiendo las indicaciones proporcionadas.

Homocedasticidad

La hipótesis de homocedasticidad implica que los errores del modelo tienen varianza constante. Dado que los errores del modelo no se observan, estudiamos dicha hipótesis sobre los residuos $e_i = Y_i - \widehat{Y}_i$, o su versión estandarizada, r_i (media 0 y varianza 1), o estudentizada (media 0, varianza 1 y distribución t de Student), $\widehat{t}_i$. En este caso vamos a calcular los residuos estandarizados que se pueden calcular con la función `rstandard`, y valoraremos la hipótesis de homocedasticidad representando gráficos de estos residuos frente a valores ajustados primero, y luego frente a cada una de las percepciones.

Calcula los residuos estandarizados r_i y representa el gráfico de dispersión de los pares $(\widehat{y}_i, r_i)$. A continuación representa para cada percepción x_j, gráficos de dispersión de los pares (x_{ij}, r_i). Observa todos los gráficos y detecta posibles patrones no aleatorios que te alertarían de desviaciones de la hipótesis de homocedasticidad.

Incorrelación

Los errores del modelo de regresión lineal múltiple que hemos ajustado se asumen incorrelados. Para verificar esta hipótesis podemos construir un gráfico de residuos (e.g. los estandarizados) frente al número de orden de cada observación (variable `empresa` en el fichero). Un patrón no aleatorio en este gráfico nos alertaría de posibles desviaciones de esta hipótesis. La impresión visual del gráfico la podemos confirmar con el test de Durbin-Watson que se puede obtener usando la función `dwtest` del paquete *lmtest*.

Estudia si la hipótesis de incorrelación es asumible para el modelo ajustado usando las herramientas anteriores.

Normalidad

Los errores de modelo se asumen normales y dicha hipótesis es esencial para desarrollar la inferencia del modelo que hemos descrito antes. Verificar esta hipótesis es por tanto crucial. Dado que el tamaño de muestra es relativamente grande ($n = 100$) un contraste de normalidad recomendable puede ser el test de Kolmogorov-Smirnov (función `ks.test`). El resultado del test lo podemos completar representando un gráfico probabilístico normal (función `qqnorm`).

Usando las herramientas anteriores estudia si la hipótesis de normalidad es asumible para el modelo ajustado. [Nota: Aplica las funciones a los residuos del modelo, por ejemplo los estandarizados.]

Linealidad

La posible falta de linealidad en la relación entre la respuesta y las variables explicativas se ha podido investigar en un primer momento observando los diagramas de dispersión

entre la respuesta y cada una de las variables explicativas. Allí se observaban pautas de relación aproximadamente lineales salvo en algunos casos donde no parecía existir mucha relación. No obstante los gráficos más adecuados para detectar posible no linealidad son los gráficos de componente más residuo que podemos obtener usando la función `crPlots` del paquete *car*.

Instala y carga el paquete *car*. Después evalúa la función `crPlots` para obtener los gráficos de linealidad escribiendo `crPlots(mod1)`. Observa el resultado.

Identificación de datos anómalos e influyentes

Datos anómalos serán aquellos que tienen un residuo asociado cuya magnitud es excesivamente grande. Para detectar estos puntos consideramos residuos estandarizados o estudentizados y localizamos valores que estén fuera del rango (-2,2), siendo muy extremos aquellos que están fuera de (-3,3).

Localiza las empresas que tienen un residuo estandarizado superior a 2.5 en valor absoluto.

Datos influyentes son aquellos que tienen un impacto desproporcionado sobre los resultados de la regresión. Si además se trata de datos aislados estos datos deberían tratarse. Para medir la influencia de las observaciones utilizamos la distancia de Cook D_i (función `cooks.distance`) y para medir el aislamiento (*leverage*) utilizamos los valores en la diagonal, de la matriz hat $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (función `hatvalues`). Medidas de influencia notablemente superiores a las del resto de observaciones corresponderían a datos influyentes (igual para el aislamiento).

Calcula los valores h_{ii} y D_i para cada empresa, evaluando las funciones `cooks.distance(mod1)` y `hatvalues(mod1)`. Representa gráficamente dichos valores frente al número de empresa (en dos diagramas de dispersión). Estudia si hay datos influyentes y/o aislados e identifica a qué cliente corresponden en su caso. Comprueba si estos datos son además anómalos en su residuo estandarizado.

El paquete *car* dispone de la función `influenceIndexPlot` que permite reproducir gráficos para hacer de una manera más sencilla esta tarea. Instala y carga el paquete y a continuación evalúa dicha función escribiendo `influenceIndexPlot(mod1)` .

Eliminación de observaciones anómalas e influyentes

Los datos anómalos y muy influyentes deben ser tratados. Cuando se tiene acceso a la fuente de los datos es necesario comprobar si corresponden a errores de medida o procesado de los datos, o bien si indican alguna situación más complicada en relación al diseño del estudio. En este caso no tenemos acceso a la fuente ni más información que la que se ha

ofrecido de modo que procederemos a eliminar dichos datos[6] Si has hecho la tarea anterior correctamente habrás podido identificar 2 observaciones anómalas y/o muy influyentes que son las empresas 7 y 100.

Elimina las 2 empresas anteriores del data frame `hatco` y vuelve a ajustar el modelo usando `lm` sobre el data frame con las 98 empresas. Al objeto resultante de evaluar la función llámalo `mod2`. Después repite los cálculos sobre la bondad del ajuste y contrastes de significación individual para este nuevo objeto. Observa qué cambios se producen con respecto a los resultados que obtuviste con los datos de las 100 empresas. En lo que sigue utiliza este nuevo data frame con 98 observaciones y el resultado del ajuste `mod2`.

Estudio de la multicolinealidad

Los contrastes de significación individual de cada percepción (x_j) nos advierten de que hay variables redundantes en el modelo. Es importante que confirmemos que esto supone un problema serio de multicolinealidad. Para descartar la existencia de multicolinealidad seguimos los siguientes pasos:

- Calculamos la matriz de correlaciones $\mathbf{R}$ entre las percepciones consideradas dos a dos. Esto se puede hacer escribiendo `R<-cor(hatco[,6:12])`. Tenemos que descartar que existan correlaciones muy elevadas.
- Calculamos el índice de condicionamiento de la matriz $\mathbf{R}$ y comprobamos que está por debajo de 30. Para obtener dicho índice puedes escribir:

```
ai<-eigen(R)$values    # autovalores de R
sqrt(max(ai)/min(ai))  # el índice IC
```

- Para cada una de las 7 percepciones calculamos el factor de inflado de varianza VIF. Todos deben estar por debajo de 10, y preferentemente por debajo de 5. Calcula dichos valores usando la función `vif` del paquete `car`. Para ello escribe `vif(mod2)`.

A partir de los resultados anteriores concluye si existe un problema de multicolinealidad en el modelo ajustado.

2.5. Selección de variables explicativas

El modelo que hemos ajustado anteriormente incorpora 7 variables explicativas. Ahora buscamos posibles simplificaciones del modelo considerando solo las variables que realmente suponen una contribución significativa a la hora de describir la variable de respuesta. Esto lo podemos hacer utilizando métodos automáticos de selección de variables como los algoritmos paso a paso (*stepwise*).

[6] Lo ideal sería hacer una eliminación gradual, uno a uno, empezando con el dato más problemático y observando en cada paso los cambios que se producen en el modelo.

A continuación vamos a aplicar un algoritmo de este tipo basado en el criterio de Akaike (AIC) para la selección de modelos. La función `step` implementa el algoritmo en el que paso a paso se permite que las variables entren y salgan atendiendo a los valores del AIC. El algoritmo termina ofreciendo el mejor modelo que será aquel que tenga el menor valor del AIC.
Con el último ajuste realizado (objeto `mod2`) escribimos:

```
> step(mod2)

Start:  AIC=281.44
fidelidad ~ velocidad + precio + flexprec + imgfabri + servconj +
    imgfvent + calidadp

            Df Sum of Sq    RSS    AIC
- velocidad  1      1.38 1472.2 279.54
- precio     1      4.63 1475.5 279.75
- imgfabri   1      6.88 1477.7 279.90
<none>                  1470.8 281.44
- calidadp   1     47.79 1518.6 282.58
- servconj   1    103.69 1574.5 286.12
- imgfvent   1    104.01 1574.8 286.14
- flexprec   1   1477.38 2948.2 347.59

Step:  AIC=279.54
fidelidad ~ precio + flexprec + imgfabri + servconj + imgfvent +
    calidadp

           Df Sum of Sq    RSS    AIC
- imgfabri  1      6.11 1478.3 277.94
- precio    1     12.43 1484.6 278.36
<none>                  1472.2 279.54
- calidadp  1     47.69 1519.9 280.66
- imgfvent  1    102.87 1575.1 284.16
- flexprec  1   1478.35 2950.6 345.67
- servconj  1   1805.21 3277.4 355.96

Step:  AIC=277.94
fidelidad ~ precio + flexprec + servconj + imgfvent + calidadp

           Df Sum of Sq    RSS    AIC
- precio    1     13.05 1491.4 276.80
<none>                  1478.3 277.94
- calidadp  1     46.45 1524.8 278.97
```

```
- imgfvent  1    165.64 1644.0 286.35
- flexprec  1   1500.56 2978.9 344.60
- servconj  1   1807.08 3285.4 354.20

Step:  AIC=276.8
fidelidad ~ flexprec + servconj + imgfvent + calidadp

           Df Sum of Sq    RSS    AIC
<none>                  1491.4 276.80
- calidadp  1     33.89 1525.3 277.01
- imgfvent  1    169.54 1660.9 285.35
- flexprec  1   2126.24 3617.6 361.64
- servconj  1   2893.83 4385.2 380.50

Call:
lm(formula = fidelidad ~ flexprec + servconj + imgfvent + calidadp,
    data = hatco)

Coefficients:
(Intercept)     flexprec     servconj     imgfvent     calidadp
   -13.3173       3.7927       7.5142       1.8377       0.4223
```

Observa que el algoritmo por defecto comienza con el modelo completo (7 variables explicativas). Nos muestra el AIC correspondiente, así como los valores que tomaría eliminando una de las variables (observa el signo - delante del nombre de la variable explicativa). Si hay alguna reducción, se pasa a la segunda iteración eliminado la variable que produce la mayor reducción del AIC. El algoritmo continua en la medida en que se observen reducciones del AIC, terminado cuando no se produzcan. En este caso el algoritmo termina seleccionando 4 variables explicativas.

El algoritmo que hemos aplicado antes es del tipo denominado *backward*, esto es, comienza con el modelo completo y paso a paso va eliminado variables. Esto corresponde al valor por defecto del argumento `direction` de la función `stepwise`. Repite el algoritmo con un procedimiento combinado *backward-forward* usando el argumento `direction='both'`. Observa el resultado.