

2

TEMA

MÓDULO:
TÉCNICAS AVANZADAS DE PREDICCIÓN

**MODELOS DE REGRESIÓN
LINEAL**



Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

Objetivos Específicos

1. Modelos de regresión lineal

- 1.1 Especificación del modelo
- 1.2 Estimación de los parámetros
- 1.3 La mejor especificación
- 1.4 La súper mejor especificación
- 1.5 Validación del modelo

Ideas clave



OBJETIVOS ESPECÍFICOS

- Identificar los pasos necesarios para la realización de una modelización.
- Conocer los diferentes métodos de resolución de los modelos lineales.
- Comparar, con rigor, diferentes modelos estadísticos.

1. MODELOS DE REGRESIÓN LINEAL

El objeto de todos los modelos estadísticos que vamos a ver a lo largo del tema consiste en:

1. **Especificar un modelo.**
2. **Estimación del modelo.**
3. **La mejor especificación.**
4. **La super mejor especificación.**
5. **Validación del modelo.**

1.1 ESPECIFICACIÓN DEL MODELO

Veamos las definiciones de los elementos que sirven para especificar un modelo.

Especificar un modelo: nuestro objetivo consiste definir la relación entre unas variables (explicativas) y una variable (explicada) a través de unas ecuaciones y unos parámetros. Podremos utilizar este modelo para explicar un comportamiento ya sucedido, o para predecir un comportamiento en el futuro.

Predictores: son nuestras variables explicativas. (edad, antigüedad, ingresos, etc.):

$$\text{Predictores} \rightarrow X = (X_1, X_2, \dots, X_p)$$

Respuesta: es nuestra variable a explicar: (compras):

$$\text{Respuesta} \rightarrow Y$$

Parámetros: son los coeficientes que expresan la influencia de las **variables explicativas** sobre la **respuesta**. Son los coeficientes desconocidos que tendremos que estimar:

$$\text{Parámetros} \rightarrow \beta = (\beta_1, \beta_2, \dots, \beta_p)$$

donde **p** es el nº de variables que tenemos para explicar **y**.

Ecuación: es la relación matemática:

$$\text{Ecuación} \rightarrow (Y)_{nx1} \approx (\hat{Y})_{nx1} = (X)_{n \times p} (\beta)_{p \times 1}$$

donde **n** es el nº de datos que tenemos y (\hat{Y}) es la **y** estimada.

o lo que es lo mismo:

$$Ecuación \rightarrow y_i \approx \hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{pki}$$

donde **i** es la observación i-ésima.

¿Qué es beta 0?

Es un parámetro que afecta a todos los registros por igual:

$$\beta_0$$

está multiplicando a un vector de unos (1). Es decir, en nuestro modelo lineal es un valor de origen para todas las observaciones de la base de datos.

¡OJO! ¿Por qué ponemos quasi-igual en lugar de = en la ecuación?

1. Porque las betas nos van a dar una aproximación a la realidad.
2. Si las estimamos bien, serán la mejor aproximación a la realidad posible con la información disponible.
3. Pero **siempre** va a haber un término que va a ser el error / perturbación.
4. El error es lo que nos equivocamos al predecir:

$$Error \rightarrow e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{pki})$$

entonces, la suma de errores al cuadrado, nos dirá como de bueno es nuestro modelo:

$$SumaResidual \rightarrow SR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



PIENSA UN MINUTO

¿Puede que no haya error en nuestro modelo?

No. Siempre hay algo de error. Siempre debe de haber algo de error. Si no tienes error, entonces:

- Tu modelo está sobre-estimado.
- Valdrá para explicar, pero no para predecir.
- No aporta nada adicional a los datos brutos que tenías en un primer momento.



Fuente www.towardsdatascience.com

Con esto, ya podemos especificar el modelo para predecir el comportamiento de los clientes:

```
# Establecemos la formula para más adelante modelizar.  
formula<-as.formula('COMPRAS~EDAD+ANTIGUEDAD+GENERO+INGRESOS+Dist_Min')  
formula  
## COMPRAS ~ EDAD + ANTIGUEDAD + GENERO + INGRESOS + Dist_Min
```

¿Por qué no otra especificación?

Primero vamos a ver cómo se estiman los parámetros de nuestro modelo y, una vez lo tengamos generado, vamos a ver los análisis correspondientes para encontrar la mejor especificación. Pero de momento, vamos a empezar con esta. Las razones:

- Por nuestra intuición.
- Por nuestro conocimiento de negocio.
- Por los análisis previos hechos.

1.2 ESTIMACIÓN DE LOS PARÁMETROS

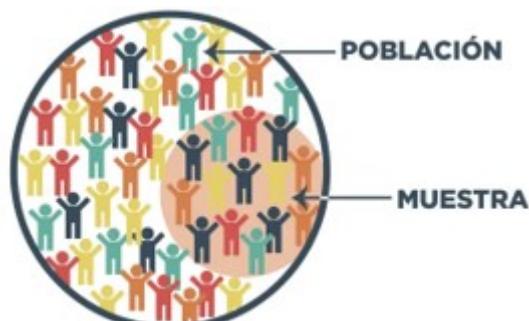
El objetivo de esta sección consiste en suministrar métodos que permitan determinar el valor de los parámetros de un modelo con precisión.



RECUERDA

Estamos trabajando con muestras y nuestro objetivo es conocer el parámetro poblacional con el estimador muestral. De nada sirve conocer el estimador para una muestra concreta, pero que falle para nuevas muestras poblacionales.

Objetivo: realizar una estimación de la función de regresión poblacional dada una muestra determinada.



Fuente www.goconqr.com

Volvemos al origen. ¿Qué es un modelo?

Ahora que sabemos de lo que estamos hablando, un modelo lo definimos como un conjunto de restricciones sobre la distribución conjunta de las variables para obtener relaciones entre las mismas. Es importantísimo centrarse en los supuestos sobre los que se construye un modelo estadístico.

¿Bajo qué supuestos basamos nuestro modelo lineal?

1. Asumimos que hay una relación lineal entre las variables explicativas y la variable explicada.
2. No puede haber colinealidad perfecta entre nuestras variables explicativas, puesto que hará que el problema no pueda resolverse.
3. Asumimos que la muestra es independiente, por lo tanto, que no hay relación entre los registros de nuestra variable respuesta.
4. Asumimos, también, que no hay relación entre los errores de mi modelo y las variables explicativas.

Nuestros estimadores serán:

- **Insesgado o centrado:** si la esperanza del estimador es igual al valor del parámetro real. Se cumplirá siempre que se cumpla la condición 4, anteriormente mencionada. Esto quiere decir que el beta estimado con nuestra muestra debe ser equivalente al beta poblacional (real), es decir, no debe contener otros efectos por detrás. Evidentemente, dependiendo del fenómeno que estemos viviendo, será más complicado poder tener betas insesgados.
- Cuanto mayor varianza de la endógena explique nuestro estimador, mejor será. Esto es cierto en términos de la SCE, sin embargo, un aspecto importante que remarcan estos modelos es la interpretabilidad y la capacidad para generar escenarios. Si nuestro modelo contiene betas sesgadas, independientemente de la predicción, podremos valorar los efectos marginales que producirá mover/actuar sobre una variable exógena, pero no se trasladará tal y como diríamos en el modelo, puesto que hay sesgo producido por otras variables no incluidas sobre las que no sabemos que tenemos que mover/actuar:

$$E[\hat{\beta}] = \beta$$

- **Consistente:** si al incrementar el tamaño muestral, el estimador se aproxima al valor del parámetro real. Se cumplirá si se cumplen las 4 condiciones de arriba. A medida que crece el tamaño muestral, el estimador converge a su parámetro real.
- Si hay homocedasticidad y no hay autocorrelación, podemos decir que los estimadores son **eficientes y óptimos**.

MÉTODO DE LOS MÍN. CUADRADOS ORDINARIOS

Consiste básicamente en minimizar la suma de cuadrados de los residuos SR:

$$\begin{aligned} SR &= e^t e = (Y - X\beta)^t (Y - X\beta) = Y^t Y - 2\beta^t X^t Y + \beta X^t X \beta \\ \frac{\partial SR}{\partial \beta} &= -2X^t Y + 2X^t X \beta = 0 \\ \beta &= (X^t X)^{-1} X^t Y \end{aligned}$$

Siendo:
 $Var(\beta) = \sigma_u^2 (X^t X)^{-1}$

Donde:
 $\sigma_u^2 = \frac{e^t e}{n - p}$

Siempre asumiendo que los residuos se distribuyen como una normal, que el error y las variables explicativas son independientes, y que los residuos son homocedásticos. En el caso de que alguno de estos supuestos no fuese verdadero, entonces la varianza de la beta no sería la anterior.

¿Qué pasa si no se cumple la condición 4?

$$E(\hat{\beta}/X) = E((X^t X)^{-1} X^t (X\beta + e)/X) = E(\beta + (X^t X)^{-1} X^t e/X) = \beta + (X^t X)^{-1} X^t E(e/X)$$

Como podemos ver en la ecuación, si $E(e/X)$ no es igual a 0, entonces la estimación de beta no sería igual a la beta real, por lo tanto, el estimador no es insesgado. Por lo que no sería eficiente. De aquí que sea tan importante las hipótesis iniciales sobre el modelo y verificar que dichas hipótesis se cumplen, de otra forma, los estimadores y, por lo tanto, las conclusiones de nuestro modelo no serían las más eficientes.

En todo este proceso, si incluimos un parámetro más (variable nueva) dentro de nuestra ecuación (especificación), dado que nos encontramos ante un problema multidimensional, todas las betas se recalcularán.

Primero, lo hacemos a mano (multiplicando matrices). Estimación de las betas de nuestro modelo:

```
X<-as.matrix(cbind(Intercept=rep(1,nrow(tabla)),dplyr::select(tabla,EDAD,ANTIGUEDAD,GENERO,INGRESOS,Dist_Min)))
Y<-as.matrix(tabla$COMPRAS)

betas_a_mano<-as.numeric(solve(t(X)%%X)%%(t(X)%%Y))
dt_a_mano<-diag(as.numeric((t(Y-X%%betas_a_mano)%%(Y-X%%betas_a_mano))/(nrow(X)-ncol(X)))%solve(t(X)%%X))**0.5
tvalue_a_mano<-betas_a_mano/dt_a_mano

tab<-as.data.frame(cbind(betas_a_mano,dt_a_mano,tvalue_a_mano))
pander(tab, split.cell = 80, split.table = Inf)
```

	betas_a_mano	dt_a_mano	tvalue_a_mano
Intercept	446.8	9.169	48.73
EDAD	1.13	0.1584	7.13
ANTIGUEDAD	-15.54	1.348	-11.53
GENERO	-28.29	5.007	-5.65
INGRESOS	0.005533	0.001279	4.327
Dist_Min	23.77	1.104	21.53

Y, ahora, utilizando la función de R **lm** (utilizando Lineal Model):

```
```{r warning=FALSE,message=FALSE}
modelo1<-lm(formula = formula,data = tabla)
pander(summary(modelo1))
```

	Estimate	Std. Error	t value	Pr(> t )
**(Intercept)**	446.8	9.169	48.73	8.617e-242
**EDAD**	1.13	0.1584	7.13	2.238e-12
**ANTIGUEDAD**	-15.54	1.348	-11.53	1.493e-28
**GENERO**	-28.29	5.007	-5.65	2.23e-08
**INGRESOS**	0.005533	0.001279	4.327	1.702e-05
**Dist_Min**	23.77	1.104	21.53	1.671e-81

Observations	Residual Std. Error	\$R^2\$	Adjusted \$R^2\$
807	65.37	0.4602	0.4568

Table: Fitting linear model: formula

## ¿Cómo interpretamos las betas?



### RECUERDA

Básicamente nuestra estimación sobre la variable respuesta va a ser igual a la suma de cada variable explicativa por su beta correspondiente. Cuando es positiva, un incremento en la variable incrementa la variable respuesta y viceversa cuando es negativa:

$$\hat{\text{Compras}} = \beta_0 + \beta_1 \text{Edad} + \beta_2 \text{Antiguedad} + \dots + \beta_6 \text{Dist. Min}$$

Lo que obtenemos con esta expresión es la estimación de las compras que va a hacer el cliente dadas unas características (variables explicativas). Esta estimación es la esperanza matemática condicionada a las variables explicativas. Por lo tanto, es una variable aleatoria cuya esperanza es el valor que estamos obteniendo.

Por decirlo de otra forma, lo que estamos obteniendo es un valor centrado de las compras que, con más posibilidad, va a hacer el cliente. Pero, una cosa que nos permite este tipo de modelos lineales es establecer intervalos de confianza (horquillas) entre los cuales es más probable que se mueva el precio.

Por ejemplo, de nada serviría que nuestro modelo de predicción nos diese un cierto nivel de compras, si el intervalo de confianza para el 90% de los casos fuese amplísimo, puesto que tendríamos un modelo con una predicción, pero con mucha inexactitud. Entonces, ¿cómo calculamos intervalos de confianza? Una parte importante es el t-value.

## ¿Cómo interpretamos el t value?

Es la medida que nos dice cuántas desviaciones estándar nuestra beta está alejada de 0 y, por lo tanto, tiene sentido tener dicha variable dentro del modelo.

El **t value** es el cociente entre la beta y el error estándar de dicha beta. Esto se distribuye como una t de student con  $(p-k+1)$  grados de libertad. El valor de dicho estadístico va asociado a la probabilidad de que dicho valor sea distinto de cero y, por lo tanto, sea significativo.

En el caso de que quisiésemos valorar si dos variables tienen la misma beta, tendríamos dos opciones:

- La primera sería ejecutar un nuevo contraste de hipótesis con la diferencia de las betas y su error estándar.
- Y la segunda opción, y la que yo os recomiendo, es realizar un nuevo modelo con la suma de ambas variables. Si el cociente de esta nueva variable (fruto de la suma de las dos anteriores a contrastar) es significativo, entonces podemos rechazar la hipótesis nula de que ambas variables tengan la misma beta.

## ¿Cómo sacamos intervalos de confianza?

Primero, vamos a sacar la predicción para una persona de 30 años, 1 año de antigüedad, género 0, 2400 euros mensuales y distancia mín. de 7:

```
```{r warning=FALSE,message=FALSE}
individuo<-as.matrix(c(1,30,1,0,2400,7))
t(as.matrix(modelo1$coefficients))%*%individuo
```
[1,]
[1,] 644.8291
```

Las compras estimadas son 644 euros. Pero, como hemos visto, esta es una estimación central de la esperanza matemática de la variable compras dadas unas condiciones o unos factores explicativos. Necesitamos saber alrededor de qué valor estamos hablando para poder tomar decisiones, o lo que es lo mismo, tenemos que preguntarnos, cómo de cierto es ese valor. ¿El cliente que hemos identificado va a hacer este gasto exactamente o va a realizar este gasto más o menos, 100 euros arriba, 100 euros abajo?

Ahora, vamos a ver el intervalo de confianza al 95% de esos 644 euros. Os hago el cálculo, manualmente, de como haríamos el intervalo de confianza de esos 644 euros. Como con todo, tenéis la función "predict" que os devuelve el intervalo de confianza ya calculado:

```
```{r warning=FALSE,message=FALSE}
bbdd<-cbind(tabla$EDAD,tabla$ANTIGUEDAD,tabla$GENERO,tabla$INGRESOS,tabla$Dist_Min)
SCR<-sum((tabla$COMPRAS-t(as.matrix(modelo1$coefficients)))%*%t(as.matrix(cbind(1,bbdd))))**2

bbdd2<-dplyr::select(tabla,COMPRAS,EDAD,ANTIGUEDAD,GENERO,INGRESOS,Dist_Min)
bbdd2$EDAD<-bbdd2$EDAD-30
bbdd2$ANTIGUEDAD<-bbdd2$ANTIGUEDAD-1
bbdd2$GENERO<-bbdd2$GENERO-0
bbdd2$INGRESOS<-bbdd2$INGRESOS-2400
bbdd2$Dist_Min<-bbdd2$Dist_Min-7

modelo_t<-lm(formula = formula,data =bbdd2)
standard_error<-summary(modelo_t)$coef[,2][1]

t(as.matrix(modelo1$coefficients))%*%individuo+2*(standard_error**2+(SCR/(nrow(tabla)-5)))**0.5
t(as.matrix(modelo1$coefficients))%*%individuo-2*(standard_error**2+(SCR/(nrow(tabla)-5)))**0.5
```
[1,]
[1,] 776.1102
[1,]
[1,] 513.548
```

Ahora, creo que nuestro modelo es mucho más potente. Ya sabemos que en el 95% de las ocasiones, este cliente se va a gastar entre 513 y 776 euros. Esto significa que nuestras decisiones para tomar acciones con este tipo de clientes tendrán que estar basadas en este último cálculo. Siendo suficientemente prudentes, podríamos coger el intervalo inferior del intervalo de confianza y, por lo tanto, solo arriesgáramos un 5% el que este valor no se cumpliese.

Por el TCL, cuantos más clientes de este tipo tengamos, la media de sus compras se va a aproximar a los

644 euros calculados, y en el 95% de los casos se moverán sus compras entre estos niveles establecidos.

### ¿Alguna forma de saber si el modelo entero es significativo?

La idea es generar una nueva hipótesis, en la que contrastemos como hipótesis nula si todas las betas, conjuntamente, son iguales a cero:

$$\frac{\frac{\sum(\hat{Y} - \bar{Y})^2}{p}}{\frac{(Y - \hat{Y})^2}{n-p-1}} \sim F_{p,n-k-1}$$

Este es otro estadístico que, en esta ocasión, se distribuirá como una F de Snedecor. Si el valor del estadístico es superior al que marca la distribución, entonces, rechazaremos que las betas de nuestro modelo son iguales que cero, por lo tanto, nuestro modelo tendrá validez. Esto funcionará cuando tenemos errores homocedásticos, en el caso de heterocedásticos no tendrá validez.



#### SABÍAS QUE...

No hay un único método de estimación de parámetros de una regresión. Entre los más utilizados están el método de los mín. s cuadrados, el método de los momentos y el método de la máx. verosimilitud. Cada uno de ellos llega a unas conclusiones de cómo estimar los coeficientes y, dependiendo del método, las propiedades de los estimadores teóricos cambian.

## MÉTODO DE MÁX. VERO SIMILITUD

El método de máx. verosimilitud es otro método de estimación. Se basa en el supuesto del tipo de distribución que sigue el término del error del modelo estadístico. A partir de esto, vamos a buscar los parámetros que hacen más probable que dichos residuos provengan de esa distribución:

$$f(u) = \frac{1}{(2\pi\sigma^2)^{0.5}} e^{-\frac{e^t e}{2\sigma^2}}$$
$$L = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{(Y-X\beta)^t(Y-X\beta)}{2\sigma^2}}$$
$$\ln(L) = -\frac{n\ln(2\pi)}{2} - \frac{n\ln(2\sigma^2)}{2} - \frac{(Y-X\beta)^t(Y-X\beta)}{2\sigma^2}$$
$$\frac{\partial \ln(L)}{\partial \beta} = -\frac{-2X^t(Y-X\beta)}{2\sigma^2} = 0$$

Siendo:  
 $Var(\beta) = \sigma_e^2 (X^t X)^{-1}$

Donde:  
 $\sigma_e^2 = \frac{e^t e}{n}$

La ecuación revela que, bajo el supuesto de normalidad del término del error, el estimador es equivalente al obtenido con el método de mín. cuadrados ordinarios, sin embargo, la varianza del estimador difiere, ligeramente, y solo coincide cuando n es suficientemente grande.

Después, analizaremos más los GLM. Por ahora, solamente para observar las diferencias con respecto a LM, os pongo los resultados que obtendríamos utilizando la función GLM que resuelve el problema de máx. verosimilitud con el método numérico de **Iterative Weighted least squares**:

```
Utilizando glm
modelo2<-glm(formula = formula,data =tabla,family=gaussian)
pander(summary(modelo2))
```

| &nbsp;          | Estimate | Std. Error | t value | Pr(> t )   |
|-----------------|----------|------------|---------|------------|
| **(Intercept)** | 446.8    | 9.169      | 48.73   | 8.617e-242 |
| **EDAD**        | 1.13     | 0.1584     | 7.13    | 2.238e-12  |
| **ANTIGUEDAD**  | -15.54   | 1.348      | -11.53  | 1.493e-28  |
| **GENERO**      | -28.29   | 5.007      | -5.65   | 2.23e-08   |
| **INGRESOS**    | 0.005533 | 0.001279   | 4.327   | 1.702e-05  |
| **Dist_Min**    | 23.77    | 1.104      | 21.53   | 1.671e-81  |

(Dispersion parameter for gaussian family taken to be 4273.44 )

```

Null deviance: 6341032 on 806 degrees of freedom
Residual deviance: 3423025 on 801 degrees of freedom

```

## 1.3 LA MEJOR ESPECIFICACIÓN

Hemos empezado con una especificación particular a criterio totalmente nuestro. ¿Era esta "fórmula" nuestra mejor especificación?



### IMPORTANTE

Es muy importante el conocimiento del negocio, de los datos y del sentido que le queremos dar al modelo, **sin embargo**, permitidme dudar que, a la primera, consigamos tener la mejor especificación del modelo.

```
```{r }  
#Especificación Inicial  
formula  
#Variables que podemos incluir en el modelo  
colnames(tabla)  
  
COMPRAS ~ EDAD + ANTIGUEDAD + GENERO + INGRESOS + Dist_Min  
[1] "COMPRAS"      "EDAD"        "ANTIGUEDAD"   "GENERO"       "INGRESOS"     "LONG"  
"LAT"            "Dist_Min"    "Dens"  
[10] "Dist_Min_h"  "Dens_h"      "Log_Ing"
```

¿Hay alguna técnica que nos ayude a identificar la mejor especificación?

Sí. Hay un conjunto de técnicas que nos ayudan a escoger un subconjunto reducido de las "p" variables que teníamos al comenzar nuestro estudio.

Primero y fundamental es conocer los indicadores más habituales para comparar modelos. Estos indicadores ofrecen una métrica de cómo es el grado de ajuste del modelo dadas un número de variables explicativas dentro de él:

1. AIC (Akaike Information Criteria):

$$AIC = \frac{SR + 2p\sigma^2}{n\sigma^2}$$

Cuanto menor sea, mejor será el ajuste de nuestro modelo. Fijaros que penaliza tanto la suma de residuos al cuadrado como el número de variables que estamos utilizando.

2. BIC (Bayesian Information Criteria):

$$BIC = \frac{SR + p\sigma^2 \ln(n)}{n}$$

Cuanto menor sea, mejor será el ajuste de nuestro modelo. Igualmente, penaliza tanto la suma de residuos al cuadrado como el número de variables que estamos utilizando.

3. R Cuadrado Ajustado/Corregido (Coeficiente de determinación):

$$R_a^2 = 1 - \frac{\frac{SR}{n-p-1}}{\frac{(Y - \bar{Y})^t(Y - \bar{Y})}{n-1}}$$

Donde:
 $\bar{Y} = \frac{\sum(Y)}{n}$

Cuanto mayor sea, mejor será el ajuste de nuestro modelo. Nuevamente, penaliza tanto la suma de residuos al cuadrado como el número de variables que estamos utilizando.

Método de selección Stepwise

El método consiste en probar distintas combinaciones de variables. Para ello, utilizamos dos métodos:

- **Método Forward:** empezamos por un modelo sin variables y vamos añadiendo una en cada paso del método.
- **Método Backward:** Empezamos por un modelo con todas las variables y vamos quitando una en cada paso del método.

Al finalizar el bucle, el algoritmo elige el mejor modelo basado en lo que nosotros queramos (BIC/AIC/R).

A continuación, paso a explicarlo paso a paso:

Step Wise: Forward	Step Wise: Backward
1.0 Generamos nuestro modelo 1 (M1). Que será el modelo sin ninguna variable explicativa	1.0 Generamos nuestro modelo 1 (M1). Que será el modelo con todas las variables explicativas
2.0 Generamos p modelos añadiendo una variable adicional al modelo anterior.	2.0 Generamos p modelos quitando una variable adicional al modelo anterior.
3.0 De los p modelos construidos cogemos el mejor (BIC/AIC/R). Será nuestro modelo (M2)	3.0 De los p modelos construidos cogemos el mejor (BIC/AIC/R). Será nuestro modelo (M2)
... repetir hasta acabar con las variables	... repetir hasta acabar con las variables
4.0 Generamos p-1 modelos añadiendo una variable adicional al modelo anterior.	4.0 Generamos p-1 modelos quitando una variable adicional al modelo anterior.
5.0 De los p-1 modelos construidos cogemos el mejor (BIC/AIC/R). Será nuestro modelo (M3)	4.0 De los p-1 modelos construidos cogemos el mejor (BIC/AIC/R). Será nuestro modelo (M3)
... así hasta que no hay más variables que añadir	... así hasta que no hay más variables que quitar
Último Paso: Elegir el mejor modelo entre los M1, ..., Mp (BIC/AIC/R)	Último Paso: Elegir el mejor modelo entre los M1, ..., Mp (BIC/AIC/R)

Para los más perfeccionistas, hay métodos que calculan el mejor modelo comparando todas las combinaciones de modelos posibles. El problema es el consumo computacional que requiere esto. Son propuestas híbridas entre forward y backward en la que, cuando se añaden nuevas variables al modelo, también se eliminan basándose en p-valores de las variables.

Vamos a utilizar la función StepAIC con nuestro ejemplo para ver la especificación que nos recomendaría. Para ello, tenemos que darle a la función, el modelo completo/vacio + todas las variables para que la función contenga todas las variables explicativas que queremos testar. El algoritmo va a tomar decisiones siempre en base al AIC del modelo:

```
formula_completa<-as.formula('COMPRAS~EDAD+ANTIGUEDAD+GENERO+INGRESOS+Dist_Min+Dens+Dis
t_Min_h+Dens_h+Log_Ing')
modelo_completo<-glm(formula =formula_completa ,data =tabla,family=gaussian)
modelo_vacio<-glm(formula =COMPRAS~1 ,data =tabla,family=gaussian)
...
```

Primero, realizamos backward. Nos sugiere eliminar la variable **ingresos** y sustituirla por el **Log_inglesos** y, además, eliminar las variables **densidad de supermercados** y **distancia mín. a un hospital**:

```
backward<-stepAIC(modelo_completo,trace=FALSE,direction="backward")
backward$anova
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
COMPRAS ~ EDAD + ANTIGUEDAD + GENERO + INGRESOS + Dist_Min +
Dens + Dist_Min_h + Dens_h + Log_Ing

Final Model:
COMPRAS ~ EDAD + ANTIGUEDAD + GENERO + Dist_Min + Dens_h + Log_Ing
```

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				797	3293571	9021.693
2	- INGRESOS	1	460.7627	798	3294032	9019.806
3	- Dist_Min_h	1	5067.7298	799	3299100	9019.047
4	- Dens	1	5608.5029	800	3304708	9018.418

En segundo lugar, realizamos **forward**. Nos sugiere quedarnos con, exactamente, la misma especificación:

```
forward<-stepAIC(modelo_vacio,trace=FALSE,direction="forward",scope=formula_completa)
forward$anova
```

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
COMPRAS ~ 1

Final Model:
COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD + GENERO + Log_Ing + Dens_h

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				806	6341032	9532.334
2	+ Dist_Min	1	1927269.64	805	4413762	9241.947
3	+ ANTIGUEDAD	1	553568.48	804	3860194	9135.801
4	+ EDAD	1	207322.54	803	3652871	9093.251
5	+ GENERO	1	149829.23	802	3503042	9061.452
6	+ Log_Ing	1	118909.67	801	3384132	9035.583
7	+ Dens_h	1	79423.87	800	3304708	9018.418

Por último, aunque ya no haría falta, vamos a probar el modelo híbrido:

```
both<-stepAIC(modelo_vacio,trace=FALSE,direction="both",scope=formula_completa)
both$anova
```

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
COMPRAS ~ 1

Final Model:
COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD + GENERO + Log_Ing + Dens_h

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				806	6341032	9532.334
2	+ Dist_Min	1	1927269.64	805	4413762	9241.947
3	+ ANTIGUEDAD	1	553568.48	804	3860194	9135.801
4	+ EDAD	1	207322.54	803	3652871	9093.251
5	+ GENERO	1	149829.23	802	3503042	9061.452
6	+ Log_Ing	1	118909.67	801	3384132	9035.583
7	+ Dens_h	1	79423.87	800	3304708	9018.418

Este tipo de algoritmos de decisión en la selección de las variables son muy útiles en la práctica. No solamente para encontrar la mejor especificación de nuestro modelo, sino para quitarnos mucho trabajo a la hora de ver qué variables entrar en el modelo final.

En nuestra base de datos, contamos con 10 variables, por lo que hacer el análisis es sencillo, relativamente.

Cuando el problema de dimensionalidad crece, por ejemplo, en bases de datos de más de 500 variables, es muy útil un simplificador en el camino que nos diga dónde centrarnos y, sobre todo, qué descartar de antemano para no invertir recursos innecesarios:



MI MODELO:

COMPRAS ~ EDAD + ANTIGUEDAD + GENERO + Dist_Min + Dens_h + Log_Ing

Fuente <https://bit.ly/33MLZ0Y>

Establecemos la fórmula para modelizar más adelante:

```
formula<-as.formula('COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD + GENERO + Log_Ing + Dens_h')
modelo_final<-glm(formula = formula,data =tabla,family=gaussian)
pander(summary(modelo_final))
```

 	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	345.3	25.76	13.4	4.285e-37
Dist_Min	20.37	1.33	15.32	1.168e-46
ANTIGUEDAD	-15.43	1.326	-11.63	5.173e-29
EDAD	1.026	0.1575	6.512	1.313e-10
GENERO	-28.15	4.923	-5.719	1.514e-08
Log_Ing	18.78	3.429	5.479	5.736e-08
Dens_h	-2.281	0.5201	-4.385	1.316e-05

(Dispersion parameter for gaussian family taken to be 4130.885)

```
-----  
Null deviance: 6341032 on 806 degrees of freedom  
Residual deviance: 3304708 on 800 degrees of freedom  
-----
```

Hemos dado un gran paso para conocer el comportamiento de nuestros clientes, pero aún queda trabajo por hacer...

1.4 LA SÚPER MEJOR ESPECIFICACIÓN

Ojo, hasta ahora, hemos contemplado el mejor modelo lineal, las mejores betas proporcionales para cada una de nuestras variables para predecir las compras.

Problemas:

(conversaciones conmigo mismo).

Pregunta: ¿Es la vida lineal?

Respuesta: Creo que no... No tiene mucho sentido pensar que el cambio en comportamiento de una persona que tiene 20 y pasa a tener 30 es igual que el de una persona que tiene 60 y pasa a tener 70.

Pregunta: ¿Podemos contemplar efectos no lineales en una regresión lineal?

Respuesta: Hay muchas maneras de contemplar no-linealidades a través de un modelo lineal.

Veamos, a continuación, algunos ejemplos:

1. Añadiendo una variable control y metiéndola (añadiéndola) en el modelo:

```
if.Edad > 30.then.Nueva.Variable = 1;  
else.Nueva.Variable = 0;
```

2. Añadiendo una variable modificada no linealmente y metiéndola (añadiéndola) en el modelo:

$$\text{Nueva.Variable} = \text{Edad}^2;$$

3. Contemplando diferentes pendientes (betas) para una misma variable. Para ello, vamos a utilizar el paquete Earth.

¿Qué hace el paquete Earth?

Desarrolla la idea de "Multivariate adaptive regression splines (MARS)". Capturamos relaciones no lineales en los datos, estableciendo puntos de corte. Los puntos de corte que nos dice MARS es donde la beta cambia para una misma variable. Lo implementamos en R y os lo explico:

El paquete nos está encontrando varias no-linealidades así es que, tenemos dos opciones:

```
modelo_final<-earth(formula = formula,data = tabla,thresh=0.1)  
summary(modelo_final)
```

Call: earth(formula=formula, data=tabla, thresh=0.1)

	coefficients
(Intercept)	596.96147
h(2.7-Dist_Min)	-47.88994
h(Dist_Min-2.7)	14.86488
h(57-EDAD)	-3.08654
h(EDAD-57)	-7.14819

Selected 5 of 5 terms, and 2 of 6 predictors
Termination condition: Rsq changed by less than 0.1 at 5 terms
Importance: Dist_Min, EDAD, ANTIGUEDAD-unused, GENERO-unused, Log_Ing-unused, Dens_h-unused
Number of terms at each degree of interaction: 1 4 (additive model)
GCV 3850.145 RSS 3038151 GRSq 0.5112197 Rsq 0.5208744

1. Emplear el modelo que nos da earth tal cual.
2. Utilizar las no linealidades para seguir manejando nuestro modelo.

¿Pero qué está haciendo el paquete Earth?

Está desmigando cada variable y partiéndola en varios trozos para ver si cada uno de los trozos de la variable explica de forma diferente y significativamente.



IMPORTANTE

Sobre todo, y muy importante. Nos fijamos en aquellas variables cuyo coeficiente tiene el mismo signo.

- **Mismo signo:** la beta cambia de tener pendiente positiva a negativa.
- **Distinto signo:** la beta pasa a ser más suave o más agresiva, pero con el mismo signo:

Las betas únicamente cambian de signo con la edad, por lo tanto, voy a llevar dicho cambio a mi modelo.

$$\begin{aligned} \text{Dist_Min} &= \beta(2.6 - x) & x < 2.6, \\ &\quad \beta(x - 2.6) & x > 2.6 \\ \text{Edad} &= \beta(57 - x) & x < 57, \\ &\quad \beta(x - 57) & x > 57 \end{aligned}$$

Podemos optar por quedarnos, puramente, con el modelo que nos proporciona "Earth" o utilizar esta técnica como una herramienta más para perfeccionar nuestro modelo GLM. Nosotros, en este caso, vamos a escoger la opción de coger los puntos de inflexión más representativos que nos da "Earth" y llevárnoslos a nuestro modelo donde tenemos control total de lo que estamos haciendo. Lo importante es sentirnos cómodos con el modelo que estamos realizando.

¿Es mejor este nuevo modelo?

```
formula<-as.formula('COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD + GENERO + Log_Ing + Dens_h')
modelo_final<-glm(formula = formula,data =tabla,family=gaussian)

tabla$EDAD_hasta_57<-((57-tabla$EDAD)<0)*0+((57-tabla$EDAD)>=0)*(57-tabla$EDAD)
tabla$EDAD_despues_57<-((tabla$EDAD-57)<0)*0+((tabla$EDAD-57)>=0)*(tabla$EDAD-57)

formula<-as.formula('COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD_hasta_57 + EDAD_despues_57 + GENERO + Log_Ing + Dens_h')
nuevo_modelo_final<-glm(formula = formula,data =tabla,family=gaussian)
pander(summary(nuevo_modelo_final))
```

 	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	445.4	22.39	19.89	7.519e-72
Dist_Min	22.05	1.128	19.54	8.18e-70
ANTIGUEDAD	-15.3	1.122	-13.64	3.088e-38
EDAD_hasta_57	-3.059	0.1752	-17.46	4.245e-58
EDAD_despues_57	-7.102	0.4737	-14.99	5.917e-45
GENERO	-27.82	4.163	-6.684	4.366e-11
Log_Ing	17.94	2.9	6.186	9.869e-10
Dens_h	-1.982	0.4402	-4.502	7.726e-06

Vamos a comprobarlo con nuestro AIC.

¿Es menor el AIC conseguido en nuestro nuevo modelo más sofisticado?

El AIC que, como veímos anteriormente, es una métrica para comparar modelos en función de la log-ver-

```
```{r}
if (AIC(nuevo_modelo_final) < AIC(modelo_final)){
 print("El nuevo modelo mejora el ajuste")
} else {
 print("El nuevo modelo no mejora el ajuste")
}```
```

```
[1] "El nuevo modelo mejora el ajuste"
```

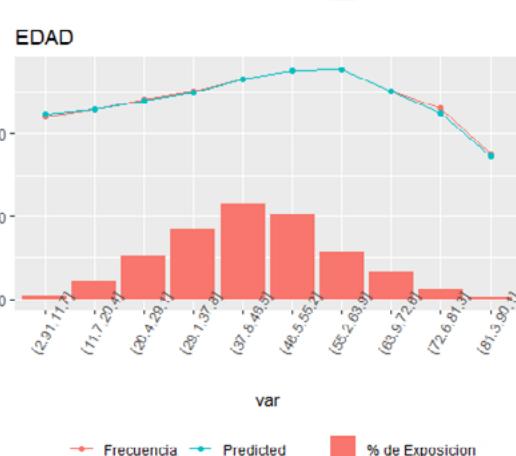
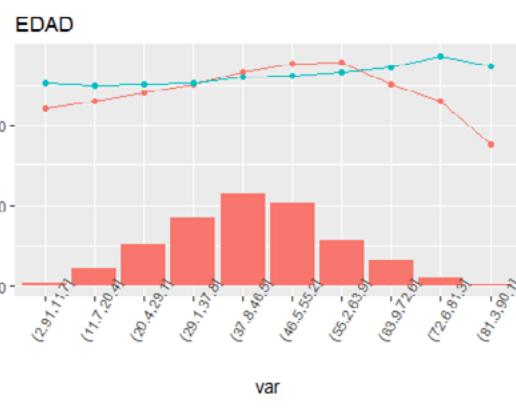
rosimilitud del mismo y las variables que utiliza para llegar a dicha estimación, nos indica que el nuevo modelo ajusta mejor a pesar de haber introducido una nueva variable en él. Hemos dividido una variable en dos partes, por lo que va a penalizar más en la métrica de AIC, sin embargo, como el ajuste es significativamente mejor, lo compensa dicha introducción.

**¿Podemos visualizar lo que estamos haciendo, por favor?**

Vamos a utilizar otra vez nuestra función **Hist** que, anteriormente, nos funcionaba para ver estructuras de datos. Ahora, la vamos a utilizar para ver el grado de ajuste de nuestros datos. En concreto, de nuestra variable **edad**:

Arriba, el anterior ajuste: Compras Reales por Edad vs Compras Estimadas por Edad. Abajo el nuevo

```
```{r fig.width = 5}
tabla1<-dplyr::select(tabla,-LONG,-LAT)
Hist(tabla1,response = tabla1[,1],predicted = predict(modelo_final,tabla1),var = tabla1[,2],n=2,breaks = 10)
Hist(tabla1,response = tabla1[,1],predicted = predict(nuevo_modelo_final,tabla1),var = tabla1[,2],n=2,breaks = 10)
```



ajuste con el modelo troceando las variables.

Parece que nos convence nuestro modelo. Hasta ahora hemos hecho lo siguiente:

1. Depurado nuestra base de datos.
2. Alimentado con nuevas variables nuestra tabla.
3. Elegido con un método de selección de variables nuestros mejores factores.
4. Desmigado no-linealidades.



MI SUPER MODELO:

**COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD_hasta_57 + EDAD_despues_57
+ GENERO + Log_Ing + Dens_h**

Fuente <https://bit.ly/33MLZ0Y>

Next step! ¿Podemos confiar en las predicciones del modelo?

1.5 VALIDACIÓN DEL MODELO

Para poder valorar si nuestro modelo puede ser utilizado para explicar y predecir faltan algunos pasos. Nos vamos a tener que adentrar en los residuos del modelo y hacer ciertas comprobaciones para asegurarnos de que todas las hipótesis de partida se cumplen.

¿Cumplen los residuos con la hipótesis de normalidad?

Tanto en el histograma que ponemos a continuación, como en el gráfico q-q plot, podemos ver cómo los residuos parecen que provienen de una distribución normal. Esto verifica que las betas que hemos obtenido están correctamente calculadas.

¿Tenemos que medir a ojo la normalidad de los residuos?

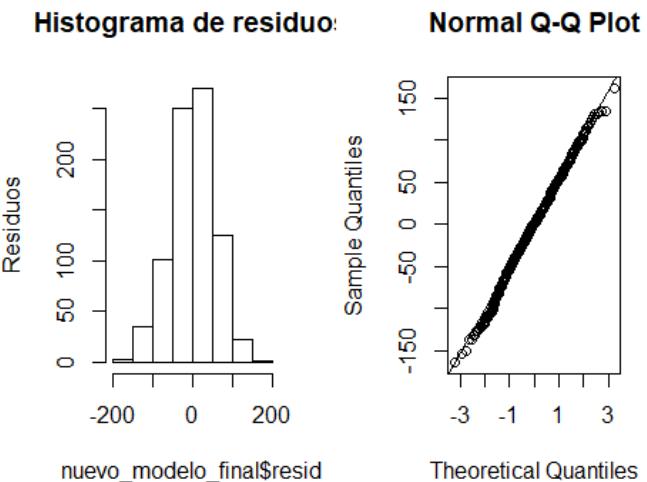
No, también hay un contraste (Jarque Bera) que dice si los residuos se distribuyen como una normal utilizando el coeficiente de asimetría de los residuos calculados y la curtosis. (Solo se utiliza para muestras grandes >500 datos):

$$JB = n \left(\frac{asimetria^2}{6} + \frac{(curtosis - 3)^2}{24} \right) \sim Chi^2$$

Donde:
 $H_0: X \sim N. la. serie. es. normal$

```
{r }
layout(matrix(c(1,2),1,2,byrow=T))
#Spend x Residuals Plot
#plot(modelo2$resid~tabla$COMPRAS[order(tabla$COMPRAS)])
#Histogram of Residuals
hist(nuevo_modelo_final$resid, main="Histograma de residuos", ylab="Residuos")
#q-qPlot
qqnorm(nuevo_modelo_final$resid)
qqline(nuevo_modelo_final$resid)
#Jarque Bera
jarqueberaTest(nuevo_modelo_final$resid)
```

```



```
#Jarque Bera
jarqueberaTest(nuevo_modelo_final$resid)
```

Title:  
 Jarque - Bera Normality Test

Test Results:  
 STATISTIC:  
 X-squared: 1.5829  
 P VALUE:  
 Asymptotic p value: 0.4532

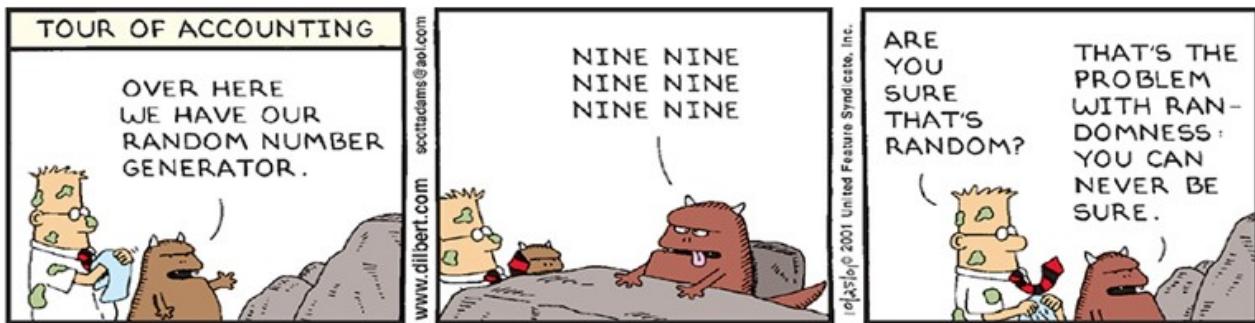
El estadístico dice que podemos aceptar  $H_0$

Así es que podemos afirmar que nuestros residuos son normales ¡Grande Jarque Bera!

¿Los residuos son independientes? [Test Autocorrelación]

Para demostrar que los residuos son independientes los unos de los otros, en el gráfico de dispersión de abajo, podemos comprobar que no presentan patrón pero, para comprobarlo matemáticamente, podemos utilizar el contraste de Durbin Watson.

¡Los errores tienen que ser aleatorios!



Fuente [www.kris-nimark.net](http://www.kris-nimark.net)

Necesitamos comprobar esto para asegurarnos de que no haya tendencia en los datos. Nuestro modelo puede ajustar de maravilla para unos datos fijos. Sin embargo, si notamos tendencia en los residuos, puede que a la hora de predecir, no acertemos ni una:

$$\begin{aligned} H_0 &: \text{los errores son independientes.} \\ H_1 &: \text{los errores no son independientes.} \end{aligned}$$

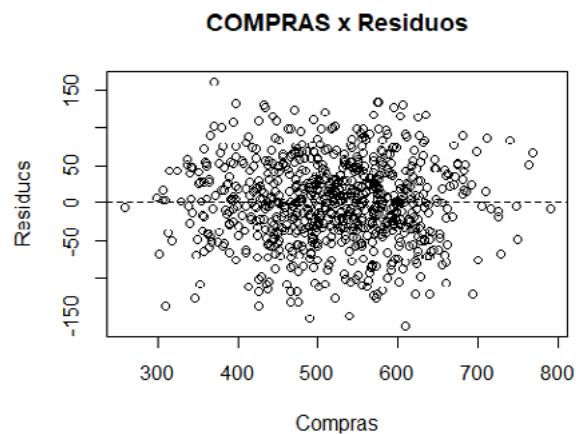
Después de aplicar el test podemos aceptar la:

$$H_0$$

Quiere decir que los errores son independientes para nuestro modelo propuesto:

```
```{r}
plot(nuevo_modelo_final$resid~tabla$COMPRAS[order(tabla$COMPRAS)],
  main="COMPRAS x Residuos",
  xlab="Compras", ylab="Residuos")
abline(h=0, lty=2)

dwtest(nuevo_modelo_final)
```
```



```
Durbin-Watson test

data: nuevo_modelo_final
DW = 1.9884, p-value = 0.4353
alternative hypothesis: true autocorrelation is greater than 0
```

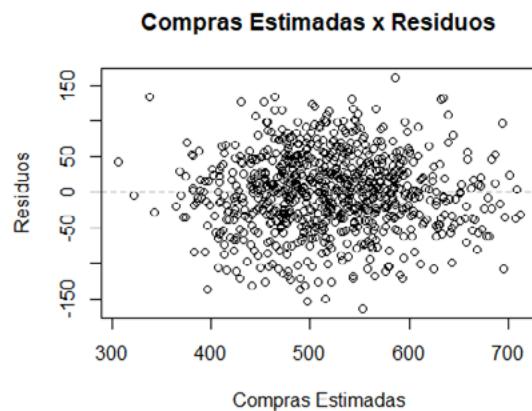
**¿Los residuos tienen varianza constante?** (Test Homocedasticidad)

Debemos chequear, también, que los errores tengan media 0 y que su varianza a lo largo de la serie es-

timada sea constante. La homocedasticidad es un criterio exigido en los modelos de regresión.

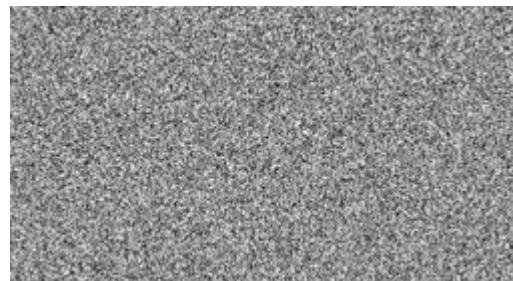
Para chequear que los errores tienen varianza constante vamos a hacer un gráfico parecido al anterior. Simplemente vamos a sustituir la variable **compras** real por la variable **compras estimadas**:

```
```{r }
plot(nuevo_modelo_final$resid~predict(nuevo_modelo_final,tabla1),
  main="Compras Estimadas x Residuos",
  xlab="Compras Estimadas", ylab="Residuos")
abline(h=0,lty=2)
```
```



El comportamiento es muy parecido, lo que pronostica que no hay heterocedasticidad. Los residuos tienen que parecer ruido blanco.

### ¿Qué es el ruido blanco?



Fuente [www.microsiervos.com](http://www.microsiervos.com)

### ¿Hay test matemático?

Of course! Breusch and Pagan lo propusieron:

$$\hat{\epsilon}_i^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + r$$

Si  $\beta_1$  y  $\beta_2$  son igual que cero, entonces los residuos no son explicativos en el modelo inicial:

$$H_0 = \text{Los Residuos son Homocedasticos}$$

$$H_1 = \text{Los Residuos NO son Homocedasticos}$$

```
```{r }
bptest(nuevo_modelo_final)
```

```
studentized Breusch-Pagan test
```

```
data: nuevo_modelo_final
BP = 7.31, df = 7, p-value = 0.3973
```

Aceptamos que los residuos son homocedásticos.

¿Qué pasaría si no conseguimos residuos homocedásticos?

Antes de pasar a ello, simplemente hacemos mención a las posibles soluciones en el caso de que el analista no consiga eliminar la heterocedasticidad en el modelo.

Modelizando, vais a encontrar, en multitud de ocasiones, modelos que presentan heterocedasticidad, es decir, que la varianza no es homogénea a lo largo de las variables explicativas. Esto es un problema en las estimaciones que hemos realizado, puesto que, como hemos comentado anteriormente, la varianza de las pendientes del modelo no estará correctamente calculada, ya que tenía hipótesis clave detrás de su cómputo que no se estarían cumpliendo. Además, las implicaciones de negocio serían las de fiarnos de un modelo que en ciertos segmentos es más incierto que en otros.

La causa más común de encontrar heterocedasticidad de los residuos en los modelos es por determinados atípicos (outliers) dentro de la base de datos que hagan que el modelo se descuadre.

Otra causa puede ser la falta de variables para la información que tenemos en la base de datos. Esto quiere decir que, cuando tenemos una cantidad exagerada de grados de libertad dentro del modelo, se incrementa la probabilidad de encontrar heterocedasticidad. Y, por último, una de las causas más comunes es una mala especificación del modelo. Quiere decir que hemos incluido variables que no debieran estar o que necesitaban una transformación previa para ser introducidas.

Tenemos dos formas de conocer la verdadera varianza de las betas que hemos obtenido. La primera es reformular el modelo de mín. cuadrados que hemos calculado; si conocemos la forma o expresión de la heterocedasticidad. En este caso, dividiremos todas las variables por la raíz de la expresión.

La segunda y más sencilla consiste en utilizar los estimadores robustos a heterocedasticidad, que son los estimadores que, aun siendo más complejos, descartan el supuesto de varianza constante y la generalizan para que sea en su forma verdadera:

$$Var(\beta/X) = n^{-1} \left(\frac{X^t X}{n} \right)^{-1} \frac{\sum xx'e^2}{n-k-1} \left(\frac{X^t X}{n} \right)^{-1}$$

Resulta interesante coger tanto la varianza basada en el supuesto de homocedasticidad como este estimador robusto propuesto para la varianza y compararlos.

¿Qué pasaría si no conseguimos residuos no autocorrelacionados?

Esto puede darse también, habitualmente, por la propia inercia temporal de los datos, por sesgos en la especificación del modelo, por la inclusión de variables retardadas de otros períodos pasados o por la utilización de medias móviles, entre otros.

Nuestra labor consiste en encontrar la forma retardada de los residuos para poder modificar el modelo de tal forma que nos dé estimadores eficientes (mín. varianza). La primera opción que tenemos es utilizar el estimador de varianza robusto que hemos visto, anteriormente, para validar la significatividad de los estimadores. La segunda consiste en modificar la ecuación original con algún tipo de retardo temporal que nos ayude a eliminar la autocorrelación. **Esta última opción es de las más utilizadas.** Si nuestro modelo está fuertemente influido por el valor que le precede, una opción que tenemos es la de realizar el modelo en diferencias con respecto al año pasado, de tal forma que nuestra nueva variable respuesta sea la diferencia entre un año y otro y no su valor original.



IDEAS CLAVE

- Adaptaciones sobre los métodos lineales como el que ofrece MARS ayudan a vencer la linealidad de las betas, convirtiendo nuestros modelos en soluciones flexibles ante casi cualquier tipo de problema.
- Los controles sobre heterocedasticidad y autocorrelación son clave para la validación del modelo y valorar su potencial en predicción.