Tema 6:

# Fórmulas y modelos

Los modelos son objetos que imitan las propiedades de los objetos reales en una forma más simple y conveniente. Con los modelos se realizan inferencias que se aplican a los objetos reales. Las diferencias entre los modelos y la realidad se denominan residuos y son una parte fundamental. Así, un mapa de carreteras es un modelo de una parte de la superficie terrestre que intenta imitar la posición relativa de las ciudades, carreteras y otros aspectos. Un buen modelo reproduce de modo preciso propiedades relevantes del objeto real al tiempo que es
sencillo de utilizar.

## 6.1. Fórmulas

Una fórmula es una expresión simbólica de la estructura de un modelo y es utilizada por las funciones de ajuste para estimar el modelo concreto. Las fórmulas fueron introducidas por Wilkinson y Rogers en 1973 y aquí se utilizan con algunas ampliaciones. Una fórmula está constituida por dos partes relacionadas: La respuesta o variable dependiente y los términos o variables independientes o predictores. Ambas partes se relacionan a través de la función **~** que, en forma de operador, permite crear un objeto de tipo **formula**. Es posible que la respuesta no esté incluida. Las expresiones que aparecen en una fórmula se interpretan como otras expresiones cualquiera excepto para los siguientes operadores:

**+ - * / : %in% ^**

cuyo significado iremos viendo a continuación.

| Expresión | Significado |
|---|---|
| T~F | T se modeliza como F |
| $F_a + F_b$ | Incluye ambos elementos |
| $F_a - F_b$ | Incluye $F_a$ excepto $F_b$ |
| $F_a * F_b$ | $F_a + F_b + F_a : F_b$ |
| $F_a / F_b$ | $F_a + F_b$ %in% $(F_a)$ |
| $F_a : F_b$ | Interacciones |
| $F_a$ %in% $F_b$ | Interacciones |
| $F \hat{} m$ | Todos los términos de $F$ cruzados hasta el orden $m$ |

La siguiente, es un ejemplo de fórmula:

**Peso ~ Altura + Edad**

que es un modelo del **Peso** como combinación lineal de la **Altura** y la **Edad**, y que representa el modelo

$$\text{Peso} = \beta_0 + \beta_1 \text{Altura} + \beta_2 \text{Edad} + \varepsilon$$

donde las tres variables son vectores numéricos.

El operador **+** permite combinar los términos. Estos no se limitan a nombres, sino que pueden ser cualquier expresión que, cuando se evalúe, se interpretará como una variable. Si queremos expresar el modelo anterior de dependencia, pero entre logaritmos de las variables, bastaría escribir

**log(Peso) ~ log(Altura) + log(Edad)**

De hecho los términos de una fórmula pueden ser:

- un vector numérico, al que corresponde un coeficiente,
- un factor, al que corresponde un coeficiente por nivel,
- una matriz, a la que corresponde un coeficiente por columna, y
- cualquier expresión que, al evaluarse, corresponda a uno de los tres tipos anteriores.

Si alguna de las variables que utilizamos es un factor, su interpretación es distinta de la de un vector numérico. Así, el modelo

**Peso ~ Altura + Sexo**

es la forma reducida de escribir el modelo

$$\text{Peso}_i = \mu + \beta \text{Altura}_i + \left\{ \begin{array}{ll} \alpha_H & \text{si} \quad \text{Sexo} = \text{H} \\ \alpha_M & \text{si} \quad \text{Sexo} = \text{M} \end{array} \right\} + \varepsilon_i$$

siendo $\alpha_H$ y $\alpha_M$ dos parámetros que representan los dos niveles de la variable **Sexo**. El modelo es equivalente a crear dos variables ficticias, por ejemplo, **Hombre**, que vale 1 en los hombres y 0 en las mujeres, y **Mujer**, con valores opuestos, y considerar el modelo

**Peso ~ Altura + Hombre + Mujer**

A menudo en este tipo de modelos no pueden determinarse todos los coeficientes de modo único. Por ejemplo, en este caso, puede modificarse arbitrariamente $\mu$ mediante un factor aditivo, sin más que restarlo a los valores de $\alpha$.

Si una variable lógica forma parte de un modelo, se considera un factor con niveles **TRUE** y

**FALSE**. Una variable de caracteres también se interpreta como un factor cuyos niveles son los diferentes valores.

También es posible utilizar una matriz, entendiéndose en ese caso que cada una de sus columnas forma parte del modelo, aunque la matriz completa se considera un solo término.

El operador **:** permite expresar interacciones entre dos o más términos, como ocurre cuando el efecto de una variable en el modelo puede ser diferente según el nivel de un factor. Los casos que se pueden dar son tres:

- \item [factor:factor]representa un término de la forma $\gamma_{ij}$ que es un conjunto de $I\times J$ constantes, una para cada pareja de combinaciones de los factores,
- \item[factor:numeric] representa un término de la forma $\beta_{j}x$ que se interpreta como una pendiente diferente para cada valor del factor, y
- \item[numeric:numeric] representa un término de la forma $\beta xy$.

Así, si las variables **Sexo** y **Provincia** pueden interactuar, la fórmula

**Sexo + Provincia + Sexo:Provincia**

expresa que se desean ajustar coeficientes para cada nivel de **Sexo** y **Provincia** y para cada una de sus combinaciones. El operador * permite escribir este tipo de fórmula más abreviadamente, así

**Sexo * Provincia**

que se expande a

**1 + Sexo + Provincia + Sexo:Provincia**

ya que **1** se incluye por defecto en cualquier fórmula, salvo que se elimine expresamente.

Si un modelo incluye términos anidados, como podría ser el caso de las variables **Comarca** y **Provincia**, puede escribirse la fórmula

**1 + Provincia + Comarca %in% Provincia**

que puede escribirse resumidamente como

**Provincia / Comarca**

El operador - se utiliza para eliminar términos, lo que permite escribir de forma más compacta un modelo complejo en el que queremos eliminar un sólo término.

Por ejemplo,

**A * B * C - A:B:C**

correspondería al modelo

**A + B + C + A:B + A:C + B:C**

y del mismo modo

**Peso ~ Altura - 1**

define un modelo sin término independiente.

La función **model.matrix** permite construir la matriz del diseño correspondiente a un modelo concreto, aunque también podemos trabajar directamente con esta matriz si el diseño es complejo y no sabemos expresarlo mediante una fórmula.

Si une una fórmula a unos datos concretos, obtendrá un modelo que puede ajustar.

## 6.2. La función lm

Esta función ajusta modelos lineales. La sintaxis es

**lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)**

cuyos argumentos son:

- **formula** es un objeto de la calse **formula** que describe el modelo que se desea ajustar.
- **data** es una hoja de datos que contiene las variables que se utilizan en el modelo. Si no se especifica o no se encuentran las variables, se toman del entorno desde el que se llama a la función.
- **subset** es un vector opcional que indica el subconjunto de datos que se utilizará.
- **weights** es un vector opcional que indica los pesos de cada variable que se utilizará en el modelo. Si no se especifica se usa mínimos cuadrados ordinarios.
- **na.action** es una función que indica qué decisión tomar cuando hay datos faltantes.
- **method** indica el método que se utilizará. Para ajuste sólo puede especificarse **qr**.
- **model**, **x**, **y**, **qr** son variables lógicas que indican si se desea o no obtener como resultado, respectivamente, el marco del modelo, la matriz del modelo, la respuesta y la descomposición QR.
- **singular.ok** es una variable lógica que indica si puede aceptarse un ajuste singular o debe considerarse un error.
- **contrasts** es una lista opcional de contrastes.
- **offset** especifica componentes conocidas a priori que deben incluirse durante el ajuste.

La función **lm** devuelve un objeto de tipo **lm** y, en el caso de respuestas múltiples, uno de tipo **c(mlm, lm)**. Es posible aplicar las funciones **summary** y **anova** a estos objetos para obtener un resumen o un análisis de varianza.